THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Understanding Spatial Dependency Structure in Urban Road Traffic Networks: Methodology and Applications in Short Term Traffic Prediction

Md. Mahmud Hasan

M.Eng., B.Sc.

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2019

School of Civil Engineering

ABSTRACT

The traffic state on a given road link depends not only on its histories but also relies on the traffic states of the adjacent links and the distant links that are not physically connected to the given road link. Estimation of traffic states of a road link necessitates an efficient use of substantial amount of data obtained from a number of road links in a large-scale network. Considering the traffic states of all links in the network to estimate the traffic states of a given target link causes significant increase in computational complexity and time due to the high dimensionality of the parameter space. In contrast, employing only the histories of the target link cannot ensure sufficient accuracy in predicting its traffic states. Therefore, it is imperative to identify the most relevant links for a given target link and utilise them in predicting the target link's future traffic states.

This study considers three approaches to identify relevant links in a large-scale urban network, namely, bivariate linear method, multivariate linear method and nonlinear method. We propose pairwise Granger causality test as the bivariate linear method, vector auto-regressive Granger causality test as well as elastic net regularisation as the multivariate linear method and regression tree as the nonlinear method for selecting relevant predictor links for the target links in the network.

This study adopts pairwise Granger causality test as the bivariate linear method to capture dependence of the traffic states of a target link to the traffic states of other links in a road network and thus identify the relevant predictor links of the target link. The Granger causality test detects the relevant predictor links that have a statistical causal effect to the target link and the magnitude of the dependence of the target link can be measured using a metric called Granger-causal strength. We propose a variable selection method that utilises this Granger causal strength to select the most significant predictor links among the relevant links of the target link. The efficiency of the proposed variable selection method is demonstrated in terms of dimensionality reduction and prediction accuracy.

The dependence of a target link on the traffic states of a road link is mediated by the presence of another road link in the network. This statistical dependence is known as conditional dependence between the traffic states of different links in the network. In the proposed bivariate linear method, the statistical dependence of the target link and each of the links in the network is evaluated separately (i.e. pairwise) by disregarding the conditional dependency among road links in the network. A multivariate linear method is proposed to capture the conditional dependence among the traffic states of the road links in the network and thus select the network-wide important links. These important links are used to provide a common set of predictors in forecasting traffic states of all the target links in the network. Using common set of predictors for all target links other than separate set of predictors for each target link reduces computational cost for the traffic prediction in a large-scale and complex road network. Based on the relevant links for each target link identified by vector autoregressive Granger causality test and Elastic net regularisation technique, a statistical approach is applied to prepare a ranking of important links that have a strong influence in the road network. Using the ranking, a common set of the most important predictor links is proposed to estimate the traffic states of any link in the network. The efficiency of the network-wide common set of predictor link is demonstrated in terms of dimensionality reduction and traffic states prediction.

The proposed bivariate and multivariate methods of selecting relevant predictor links assume linear dependence of traffic states among the road links in the network. However, the dependence can be nonlinear in real traffic condition. This study proposes the use of regression tree to capture the nonlinear dependence among traffic states of the road links in the network and detects a set of relevant links for each target link. The relevant predictor links selected by this regression tree-based nonlinear method are compared with those selected by the linear method based on the vector autoregressive Granger causality test to assess their capabilities to reduce the dimensionality of the predictor sets and improve the accuracy of predicting traffic states in short term basis.

The urban road network of Brisbane in Australia is selected as a test bed and 5-minute interval traffic flow data are used to demonstrate the application of the proposed variable selection methods in building short-term traffic prediction models. The case studies show that the

proposed methods are effective in detecting spatial dependence among road links in the network and selecting a parsimonious set of relevant predictor links for an individual target link or a whole road network. The parsimonious set of relevant predictor links can then be used as the input variables in short-term traffic prediction models, which have reduced computational complexity while ensuring higher prediction accuracy.

DECLARATION BY AUTHOR

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

……

Md. Mahmud Hasan
School of Civil Engineering,
The University of Queensland

Publications included in this thesis

No publications included

Submitted manuscripts included in this thesis

1. Chapter 2

Hasan, M. M., Kim, J., Prato, C., Identifying spatial dependence of traffic states in urban road networks by using Granger causality, submitted to *Transportmetrica A: Transport Science.*

Details of the contribution:

- conception and design of the study
- analysis and interpretation of the research data
- drafting all parts of the publication

2. Chapter 3

Hasan, M. M., Kim, J., Prato, C., Identification of the most influential road links in urban traffic networks, submitted to *IEEE Transactions on Intelligent Transportation Systems.*

Details of the contribution:

- conception and design of the study
- analysis and interpretation of the research data
- drafting all parts of the publication

Other publications during candidature (Peer-reviewed conference papers):

1. Hasan, M. M., Kim, J., Prato, C., 2018. Identifying relevant predictors and spatial relations of traffic parameters for network-wide short-term traffic prediction, *In 97th Transportation Research Board (TRB) Annual Meeting*, Washington D.C., United States, 7-11 January.

2. Hasan, M. M., Kim, J., Prato, C., 2017. Spatial variable selection methods for network-wide short-term traffic prediction, *In 39th Australasian Transport Research Forum (ATRF) Proceedings*, Auckland, New Zealand, 27-29 November.

3. Hasan, M. M., Kim, J., 2017. Granger causality method to detect spatial dependency in a road traffic network and its application in traffic flow prediction, *In 96th Transportation Research Board (TRB) Annual Meeting*, Washington D.C., United States, 8-12 January.

4. Hasan, M.M., Kim, J., 2016. Analysing functional connectivity and causal dependence in road traffic networks with Granger causality, *In 38th Australasian Transport Research forum (ATRF) Proceedings*, Melbourne, Australia, 16-18 November.

Contributions by others to the thesis

No contributions by others.

Statement of parts of the thesis submitted to qualify for the award of another degree

No works submitted towards another degree have been included in this thesis.

Research Involving Human or Animal Subjects

No animal or human subjects were involved in this research.

ACKNOWLEDGEMENTS

First of all, author would like to thank the Almighty for the successful completion of this research.

The author expresses heartfelt gratitude to a number of people who played a significant role while conducting the research.

The author is greatly indebted to his supervisors Dr. Jiwon Kim and Prof. Carlo Prato for their invaluable guidance, support and encouragement throughout this journey.

The author would like to convey his deep appreciation to the committee chair, Prof. Mark Hickman and committee members, Dr. Mehmet Yildirimoglu and Dr. SangHyung Ahn for their constructive feedback on this research.

The author would like to acknowledge Department of Transport and Main Roads in Brisbane, Queensland for providing necessary data to perform the case studies of this research.

The author extends sincere gratitude and thankfulness to the UQ transport group and all his friends in the University of Queensland for their help and co-operation.

Finally, the author is indebted to his parents Humayun Kabir and Nurun Nahar, without whom this study would have been impossible. The author is also grateful to his spouse Nazmun Nahar and his siblings Azmin Nahar, Shahanoor Alam and Mehedi Hasan for providing emotional strength and constant support to achieve his dream of being a doctorate.

The author would like to dedicate this thesis to his parents, spouse and siblings.

Financial support

This research was supported by an Australian Government Research Training Scheme Scholarship.

Keywords

Time series analysis, spatial dependence, variable selection, Granger causality, Granger causal strength, elastic net regularisation, regression tree, important predictors, nonlinear analysis, traffic prediction.

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 090507, Transport Engineering, 80%

ANZSRC code: 090599, Civil Engineering not elsewhere classified, 20%

Fields of Research (FoR) Classification

FoR code: 0905, Civil Engineering, 80%

FoR code: 0999, Other Engineering, 20%

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

ADF	Augmented Dickey Fuller
AIC	Akaike Information Criterion
AR	Auto Regression
ARIMA	Autoregressive Integrated Moving Average
BIC	Bayesian Information Criterion
DTMR	Department of Transport and Main Roads
EN	Elastic Net
GC	Granger causality
ITS	Intelligent Transport Systems
LASSO	Least Absolute Shrinkage and Selection Operator
MA	Moving Average
MAE	Mean Absolute Error
MSE	Mean Squares Error
PTDS	Public Traffic Data System
RESET	Regression Equation Specification Error Test
RMSE	Root Mean Square Error
RSS	Residual Sum of Squares
VAR	Vector Auto Regression

Chapter 1

INTRODUCTION

1.1 Background

A reliable and accurate measure of traffic states is vital for real time traffic control and management. Traffic managers can take necessary actions to ensure enhanced road network performance and provide accurate travel information to the road users when correct estimations of future traffic states are available. A critical part of estimating future traffic states is the efficient utilisation of ample data resources that are available to traffic authority. Using all available data in traffic prediction increases prediction accuracy marginally compared to the substantial escalation in computation costs. Exploring the most relevant data is necessary to maximise the accuracy of the prediction by minimising computational costs. One way to obtain the most relevant data for traffic prediction is to identify the proper set of input predictor variables.

Determining the proper set of input predictors to forecast the traffic states of a given target link in a large-scale road network is a daunting task. The traffic states of a given target link depend not only on its histories but also the traffic states of the other links in the network. It is impractical to use only the histories of the target link itself or the traffic states of all the links in the network as the input variables in traffic prediction. Therefore, finding a systematic approach to detect the relevant predictor links for a target link is imperative for traffic prediction in a large-scale road network. The goal of this approach would be to consider only a subset of all available predictor links as the input predictor variables so that we can build a parsimonious and simpler prediction model while achieving greater accuracy in the traffic prediction.

In developing a systematic approach to selecting input variables for traffic prediction models, it is important to understand the spatial relation among the road links in the network and its implementation to select the most relevant predictor links for a given target link. The spatial relation refers to the statistical dependence between two road links in terms of their observed

traffic states, regardless of their structural connectivity or physical connection. As such, two distant links characterized by low structural connectivity can show significant spatial relationship when traffic flow time series from these two links exhibit high statistical dependency. Once the spatial relations among road links are properly captured, this information can be used to enhance the prediction of the traffic states of a target link by taking into account the traffic states of its spatially relevant links.

The trend of utilising the spatial relation of road links in traffic prediction is prevalent in recent years. The earlier researches in traffic prediction only considered the historical past of the target link in predicting the future traffic states of that target link (Ahmed and Cook 1979; Okutani and Stephanedes 1984). A number of researchers recognised the significance of considering the histories of the adjacent links of the target link in traffic prediction. They utilised few road links in a fixed spatial boundary only limited to the nearest upstream link (Hobeika and Kim 1994; Stathopoulos and Karlaftis 2003; Sun et al. 2006), nearest upstream and downstream links (Chandra and Al-Deek 2009; Kamarianakis et al. 2010; Pascale and Nicoli 2011) and nearer neighbour links (Kamarianakis and Prastacos 2003, 2004; Zhang and Ren 2013) as the relevant predictor links in predicting the traffic states of the target link. However, the road links that are not in close proximity to the target link were largely neglected in identifying the spatially relevant predictor links for the target link. Although some recent studies utilised the distant locations in selecting the relevant predictor links for the target link (Li et al. 2015; Xu et al., 2016; Yang et al. 2017), these studies mostly considered a few target and predictor links in a small road network or a freeway as the case study. However, the actual challenge appears in developing a method to identify the spatially relevant predictor links for the target link in a large-scale and complex urban road network. As these existing studies used a small road network, therefore, a systematic approach to select a parsimonious set of relevant predictor links for a target link could not be established. Furthermore, using separate relevant predictor set for each target link, as demonstrated in existing studies, increases computational complexity and time in forecasting traffic states of all the target links in a large-scale urban network. Proposing a method of identifying a common set of relevant predictor for all the target links in a large-scale network can reduce the computational cost in traffic prediction.

The spatial relation between road links in the network can be identified by two types of approaches, namely, bivariate analysis method and multivariate analysis method. In the bivariate analysis method of spatial relation, the statistical dependence of the target link and each of the links in the network is evaluated separately (i.e. pairwise) by disregarding the conditional dependency among road links in the network. Conditional dependency is the statistical dependence of the target link on a link that is mediated by the presence of another link in the network. The multivariate analysis method of spatial relation considers the conditional dependency among road links in the network. In this method, the statistical dependence of the target link on a link is evaluated by considering the effects of all other links in the network. Another aspect to consider is whether the spatial relation between road links is described as a linear model or a nonlinear model. In the linear modelling approach, the statistical dependence of the target link on other links is assumed to be linear. In real traffic conditions, the spatial relation can be nonlinear. Therefore, the nonlinear modelling approach is useful to capture the actual spatial relation among the road links in the network.

Based on these considerations, this study explores three different approaches to identify spatial dependency structure among road links in the network, namely, (i) bivariate linear method, (ii) multivariate linear method and (iii) nonlinear method. The bivariate linear method employs *pairwise Granger causality test*; the multivariate linear method employs *vector autoregressive Granger causality test* as well as *elastic net regularisation*; and the nonlinear method employs the *regression tree* method. These methods can select a distinct set of relevant predictors for each target link in the road network. The Granger causality test identifies directed functional or causal interactions of different variables in time series data. The Granger causality test has become an established method for analysing statistical causal relationship in many research domains such as neuroscience and economics. However, it has not been widely explored in the area of traffic research. Elastic net is a regularisation technique which prevents statistical over-fitting for a predictive model by using a penalty term. In a linear regression analysis, the elastic net regularisation can remove irrelevant or redundant predictors from the model. The regression tree is a form of nonparametric decision tree approach which explores the structure in the dataset without assuming an underlying distribution. Regression tree can handle the nonlinear and complex interactions of numerous features within the dataset. Regression tree segregates data

into smaller, non-overlapping and homogenous subsets to detect the interactions of the target and predictor variables.

1.2 Research questions

In order to propose the methodology of determining spatial dependency structure in urban road networks, the following four research questions were formulated, which can serve as a guideline for the scope of this thesis and give a direction to establish the research objectives.

The four research questions addressed in this thesis are:

1. How are traffic parameters of different road links spatially related to each other in an urban road network? Can the dependency relations be modelled as a linear model? How can this spatial dependency be quantified?
2. Does the spatial relation structure identify a set of important road links for a given link or the whole network?
3. Are the relationships among traffic parameters of different road links nonlinear? If so, how can the methods to address Research Questions 1 and 2 be extended to capture the nonlinear relationship of traffic parameters?
4. How is short-term traffic prediction related to the understanding of the spatial dependency structure of the network? Can the knowledge of the spatial relation structure lead to a parsimonious traffic prediction model with better prediction accuracy?

1.3 Research objectives

This study aims to develop systematic approaches to select input variables for short-term traffic prediction models by discovering the spatial relationships among road links in a large-scale urban network and identifying a reduced set of important predictors. We propose variable selection methods that can identify the most relevant predictor links for an individual road link as well as a group of important predictors for the whole network by using bivariate linear method, multivariate linear method and nonlinear method.

The objectives of this study are:

1. Propose statistical methods to detect the spatial dependency structure among the road links and identify a set of important predictor variables for an individual link's traffic prediction.
2. Develop a framework to identify a common set of most important road links in the network and its application in network-wide traffic prediction.
3. Propose a nonlinear extension of the variable selection method to capture nonlinearity in the spatial dependency structure among the road links and further enhance the identification of important predictors for traffic prediction models.

1.4 Research framework and stages

This research work can be divided into three stages: 1) proposing a bivariate linear method of selecting the relevant predictor links for a given link 2) developing a framework using a multivariate linear method of identifying the important predictor links for the whole road network and 3) proposing a nonlinear method of selecting relevant predictor links for a given link. The details methodology and application of these three stages are included in the chapter 2, 3 and 4 respectively. This thesis has been written in the format of 'Thesis by publication'. Therefore, instead of formulating one single chapter containing all the relevant literature reviews, separate sections have been provided in the relevant chapters. This approach has been adopted to ensure a reader is able to relate the literature review with its relevant methodologies adopted for the specific stages of research. The relation between the research elements (research questions and research objectives) in each of three stages of the research is described below:

In stage 1 of the study, the proposed bivariate linear method based on pairwise Granger causality test measures the statistical dependence between the target link and each of the links in the network separately. The strength of the statistical dependence of the target link to other links is quantified by using the Granger causal strength metric. A hierarchy or ranking of the predictor links is then established using the Granger causal strength possessed by each predictor link. Based on the ranking, the most relevant predictor links for the target link are selected as the input variable set for the prediction of the target link. The performance of short-term traffic prediction

model using the most relevant predictors is compared with the performance of other existing traffic prediction models. Therefore, objective 1 of the research can be accomplished by exploring the answer of research question 1 and 4 through the proposed method in stage 1. The relation between the research elements in stage 1 is illustrated in Figure 1.1.

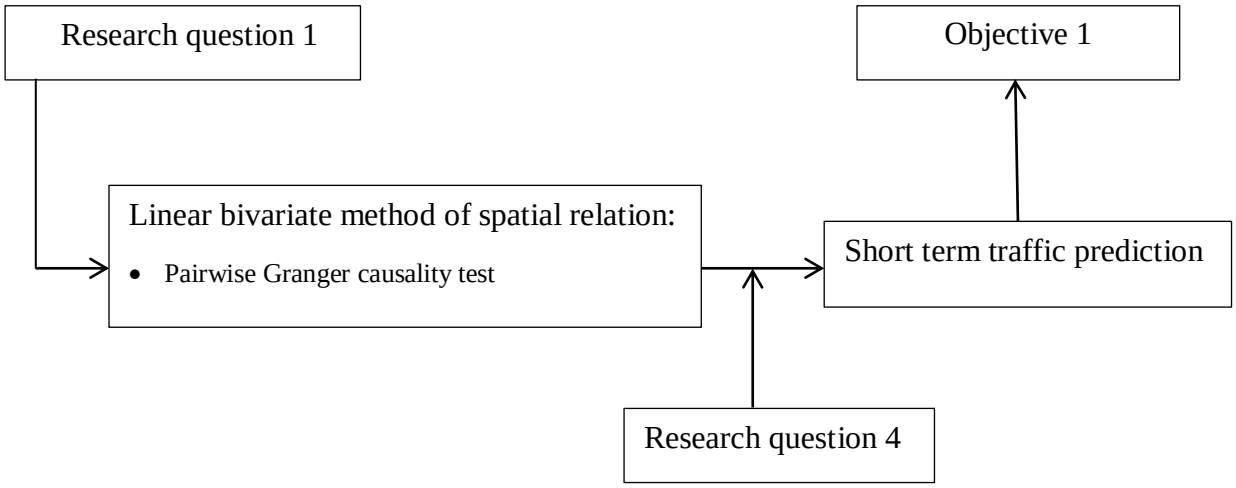


Figure 1.1: Connection between research elements in stage 1.

In stage 2, the proposed multivariate linear method based on vector autoregressive Granger causality test and elastic net regularisation measures statistical dependence between the target link and the other links in the network. Initially, all the road links in the network are considered as the possible predictor links of the target link and then a set of relevant predictor links for each of the target links is identified by vector autoregressive Granger causality test and elastic net regularisation separately. Then the importance of each predictor link in the road network is evaluated by the number of times the predictor link is selected as the relevant predictor link in the network. The predictor links are ranked according to their importance and a set of the most important predictors are identified based on the ranking. This parsimonious set of the most important predictors is the common set of predictors for all target links in the network. A short term traffic prediction model for each target link in the network is developed by utilising the common set of predictors as the input variables. It is practically easier to use a common set of predictors in predicting traffic states of all target links in the large scale network rather than to use a separate set of predictors for each target link. The performance of traffic prediction based

on the common set of the most important predictors is also compared with other existing traffic prediction models. Thus objective 2 can be fulfilled by solving the research question 1 and 4 via the proposed method in stage 2 of the research. Figure 2.2 depicts the connection between research elements in stage 2.

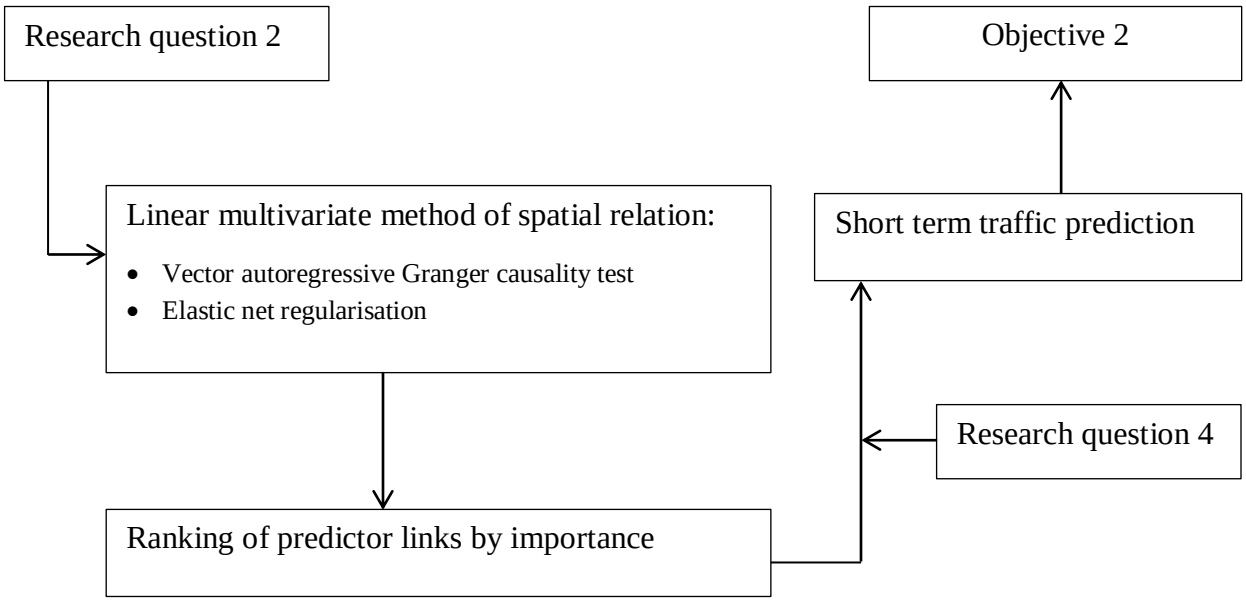


Figure 1.2: Connection between research elements in stage 2.

In stage 3 of the study, the proposed nonlinear method based on regression tree measures spatial dependence between the target link and the other links in the network. This method initially considers all the links as the possible predictor links for each target link and then detects a set of distinct relevant predictor links for each target link. As real traffic situation is a nonlinear process, the relevant predictor links selected by a nonlinear method should be more effective input variables for traffic prediction than the relevant predictor links selected by a linear method. The performance of the relevant predictor links selected by regression tree is compared with the relevant predictor links selected by vector autoregressive Granger causality test in terms of short-term traffic prediction. Hence objective 3 can be achieved by answering research question 3 and 4 through the proposed method in stage 2 of the research. Figure 3.3 shows the relation between research elements in stage 3.

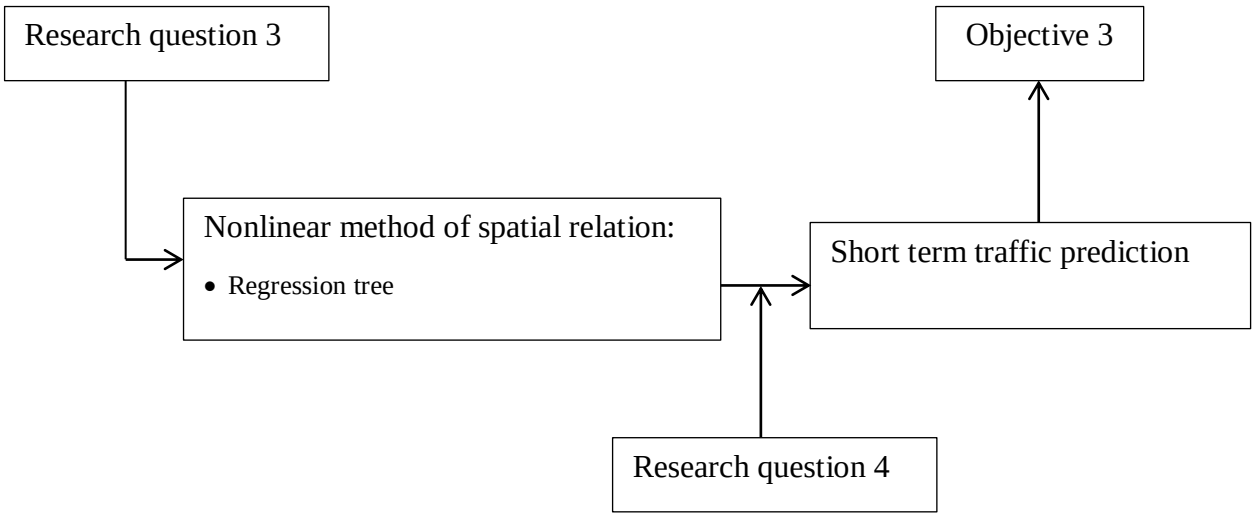


Figure 1.3: Connection between research elements in stage 3.

1.5 Significance of the research

The study presents a systematic approach of selecting input variables for traffic prediction in a large-scale urban road network, which consists of hundreds of road links. The proposed methods identify the traffic states of the most relevant predictor links as input variables for predicting the traffic states of an individual target link by using spatial relation of the links in the network. This parsimonious set of input variables ensures higher prediction accuracy and minimises computational costs in real time traffic prediction. This outcome helps traffic authorities to forecast future traffic condition, take prompt decision if required and advise road users accordingly. The proposed methods can be utilised as effective tools for traffic management and control scheme. The study also proposes method that identifies the network-wide most important links and selects a set of common predictor links for all of the target links in the network. Utilising the proposed method, traffic authorities can have the capability to monitor and predict traffic states with a limited expenditure in the number of sensors.

1.6 Outline of the thesis

The thesis is organised as follows:

Chapter 1 introduces the background of this research, establishes the research questions, research objectives and research framework, and describes the outline of the research. Chapter 2 proposes a statistical method of selecting the most relevant predictor links for individual target link using bivariate linear Granger causality test. The implementation of the proposed method in short term traffic prediction and evaluation of its prediction performance are also discussed in this chapter. Chapter 3 presents a framework to identify a common set of the most important predictors for the whole network using multivariate linear method based on vector autoregressive Granger causality and elastic net regularisation. A comparison of the prediction performance of these two methods is also demonstrated. Chapter 4 describes a nonlinear method of relevant predictor selection using regression tree and compares the efficiency of the linear and nonlinear methods of relevant predictor selection in regards to short term traffic prediction. Chapter 5 summarises the conclusions and the findings of this study and also proposes future research directions.

Chapter 2

IDENTIFYING SPATIAL DEPENDENCE OF TRAFFIC STATES IN URBAN ROAD NETWORKS BY USING GRANGER CAUSALITY

2.1 Introduction

For real-time traffic control and management, it is crucial to obtain accurate estimates and predictions of the traffic states on a road network. Accurate information on current and future traffic states allows traffic managers to take effective actions to smooth the network flow and road users to choose routes to avoid congestion proactively. A critical part of traffic estimation and prediction is building models by selecting a proper set of input and output variables. Given a large number of road links in a network, determining input variables for a prediction model can be a challenging task due to the potential trade-off between model accuracy and computational complexity. For instance, the prediction of the traffic state of a target link could consider models ranging from the simplest form, which only includes the past states of the target link itself as input variables, to the most complex form, which includes all the other links in the network as input variables. Adding input variables to a model may not only improve the model accuracy to some extent, but also increase the computational complexity and hence cause issues in applying the model in the real-time prediction context. Consequently, there is a need for a systematic approach to determining appropriate predictors for a target traffic link state. Finding the spatial relations among the road links in the network can be a solution that allows to identify the most relevant road links for a given target link.

Earlier studies on traffic prediction tended to build a model based only on the past states of the target link, without taking into account the effect of the surrounding traffic (Ahmed and Cook, 1979; Okutani and Stephanedes, 1984; Smith et al., 2002; Clark, 2003). Later, researchers recognised the importance of considering adjacent links in the prediction model. For instance, studies identified that utilising the data of the adjacent upstream link is useful to compute the approximate traffic states of that road link (Hobeika and Kim, 1994; Stathopoulos and Karlaftis,

2003; Sun et al., 2006). Some studies revealed that the traffic condition of both upstream and downstream links have a significant effect on the traffic condition of a given link (Chandra and Al-Deek, 2009; Kamarianakis et al., 2010; Pascale and Nicoli, 2011). Another stream of research focused on selecting the effective neighbourhood or neighbouring links for the given link. A hierarchical system of two orders of upstream neighbour links was defined as the spatial boundary in an urban road network and was proven to be an improvement over considering only first order neighbours to obtain the future information of a link (Kamarianakis and Prastacos, 2004). Notably, higher orders of neighbour links were not considered. Also, a method was proposed to identify the spatio-temporal correlation of road links based on the distance that can be covered by a vehicle within a fixed travel time (Min and Wyntar, 2011). Notably, this method could not uncover a relation between a given road link and its downstream links. More recently, a hybrid model based on Granger causality test and Bayesian network was proposed to identify spatio-temporal causal relation of road links that are directly connected to each other (Zhang and Ren, 2013). Ermagun and Levinson (2018) discussed the shortcomings of using predefined locations of predictor links (e.g. neighbour links) or considering similar spatial effects in predicting traffic state of each given target link. They mentioned that predefined predictor links can increase error in traffic prediction if it is not spatially relevant to the target link.

Recent studies have attempted to understand the dependence between road links that are not physically connected. Incidents on the road network were showed to be influenced by most of the links in that network (Chandra and Al-Deek, 2009). Then, the dependence of road links in a large road network was analysed by using a two-step approach based on LASSO (Least Absolute Shrinkage and Selection Operator) regression and Granger causality (Li et al., 2015). The results showed that the time series of a road link can be dependent on the time series of other road network links, even if they are far apart. Notably, the two-step approach considered a fixed number of time series to be selected via LASSO regression before applying the Granger causality test to further reduce their number, although some time series that are omitted by LASSO regression may have significant Granger-causal relations to the given time series. Lastly, the traffic states of a given link were found to relate to not only its own history and nearest road links but also other road links that may not be physically connected with the given link (Xu et al., 2016; Yang et al., 2015, Pavlyuk, 2019).

It is evident from the review of previous literature that a research gap exists in utilising the spatial relationship among the traffic states of road links for short term traffic prediction. Previous studies primarily considered the effect of the traffic states of neighbouring links on the traffic states of a given link and thus ignored the influence of the traffic states of other distant road links in the network. Moreover, these studies selected some predefined neighbour links as the relevant predictors to forecast the traffic states of a given link rather than proposing a systematic approach of predictor selection. Methods used in previous studies are not efficient in identifying the most relevant predictors set among links in the network to improve the prediction accuracy of the prediction model.

To address this gap, this study aims to develop an enhanced input variable selection method for traffic prediction models. As it is expected that the traffic state of a road link is related to the traffic states of not only the adjacent links, but also the road links in the network that are not directly connected to the target link, this study contributes to the literature by proposing a statistical method to measure spatial dependence between road links and identify good predictor links of the traffic state of the target link. The proposed approach initially considers all road links in the network as possible predictors of a given target link and then selects the relevant ones by means of the *Granger causality* test (Granger, 1969; 1980). Moreover, this approach quantifies the strength of the causal relation by using the 'Granger-causal strength' metric (Geweke, 1984; Bressler and Seth, 2011), in order to rank the predictor links according to their importance. Granger causality test is a well-established statistical approach in neuroscience and economics. In neuroscience, Granger causality test was employed to identify the directed interactions among different regions of brain network (Bernasconi and konig, 1999; Goeb et al., 2003; Hamilton et al., 2011; Dhamala et al., 2008). These studies concluded that Granger causality test is an effective tool to detect the causal dependence among different parts of neural system and it provides a valuable insight into brain functioning.

Our proposed method is intended to serve the input variable selection stage for a traffic prediction model, where a set of road links should be selected as model input such that input predictor links are informative in predicting the traffic state of the target link. Accordingly, the method helps improve the prediction accuracy while keeping the number of variables to a

minimum to obtain a parsimonious model. It should be noted that the focus of this study is on the development of an 'input variable selection method' for a short-term traffic prediction model, not the traffic prediction model itself.

The remainder of the chapter is organised as follows. Section 2.2 describes the spatial variable selection method, its implementation and performance evaluation. Section 2.3 illustrates the case study and details the application of the proposed method to the large-scale road network. Section 2.4 illustrates and discusses the results of the case study before section 2.5 summarises the conclusions of this study and proposes future research avenues.

2.2 Methodology

We propose a spatial variable selection method for a target link on a road network that consists of (i) selection of the predictors that cause the traffic state of the target link, (ii) measurement of the causal effect strength, and (iii) ranking of the selected predictors.

2.2.1 Granger causality method for the selection of predictors

The selection of the predictors is based on the Granger causality (GC) test. The GC test is a statistical test for prediction of causality between two time series, and it is based on two major principles: (i) the cause happens prior to the effect and (ii) the cause makes unique changes in the effect (Granger, 1969; 1980). A time series x_t is said to *Granger-cause* another time series y_t if adding the past values of x_t can improve the predictions of y_t. Mathematically, the GC test involves the comparison of the following two linear time series models:

$$y_t = a_0 + \sum_{j=1}^{P} a_j y_{t-j} + \varepsilon_t \tag{2.1}$$

$$y_t = a'_0 + \sum_{j=1}^{P} a'_j y_{t-j} + \sum_{j=1}^{P} b_j x_{t-j} + \varepsilon'_t \tag{2.2}$$

where $a_j, a'_j,$ and b_j are regression parameters, a_0 and a'_0 are intercepts, and ε_t and ε'_t are the residuals (or prediction errors), *j=1,2,3,…,P* and *P* is the optimal time lag. The GC test compares

the prediction errors of models (2.1) and (2.2) to measure the effect of adding x_{t-j} for predicting y_t. If model (2.2) produces better predictions of y_t than model (2.1) does, namely ε_t' is statistically significantly less than ε_t according to a F-test, then time series x_t is considered informative in predicting time series y_t and x_t is said to *Granger-cause* y_t (Granger, 1969).

An important condition required in the GC test is that the time series variables are stationary. Accordingly, each time series variable needs to be evaluated by a *unit root test* such as the Augmented Dickey Fuller (ADF) test (for details, see Dickey and Fuller, 1979). The null hypothesis is that a time series variable is non-stationary and has a unit root: if the null hypothesis is not rejected, the time series variable is considered non-stationary and needs to be *detrended* to make the series stationary.

We propose the application of the GC test to determine a potential causal relation between two time series of traffic state variables. More specifically, given a target link m ($m = 1, \cdots, N$), where N is the number of links in the network, we determine whether a time series variable from another link n ($n = 1, \cdots, N$ and $n \neq m$) helps the prediction of the traffic state on link m by performing the bivariate GC test. Let $\theta_{m,t}$ denote a variable representing the traffic state on link m at time t, then the bivariate GC test between target link m and predictor link n can be performed via the following linear equation:

$$\theta_{m,t} = c_{m,n} + \left(\pi_{m,m}^{(1)}\,\theta_{m,t-1} + \pi_{m,m}^{(2)}\,\theta_{m,t-2} + \cdots + \pi_{m,m}^{(p)}\,\theta_{m,t-p}\right) + \left(\pi_{m,n}^{(1)}\,\theta_{n,t-1} + \pi_{m,n}^{(2)}\,\theta_{n,t-2} + \cdots + \pi_{m,n}^{(p)}\,\theta_{n,t-p}\right) + \varepsilon_{m,n} \quad (2.3)$$

where $\pi_{m,m}^{(p)}$ is the coefficient capturing the relation between traffic states at time t and $t-p$ for the same link m, $\pi_{m,n}^{(p)}$ is the coefficient capturing the relation between the traffic state on target link m at time t and the traffic state on predictor link n at time $t-p$, and $c_{m,n}$ and $\varepsilon_{m,n}$ represent respectively the intercept and prediction error. The GC test consists in testing the null hypothesis $\mathrm{H}_0: \pi_{m,n}^{(1)} = \pi_{m,n}^{(2)} = \cdots = \pi_{m,n}^{(P)} = 0$.

We determine the optimal time lag order P to balance good fit with parsimony by comparing model selection criteria namely the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Ding et al., 2006; Seth, 2010; Bressler and Seth, 2011). At first, the bivariate regression model is fitted with lag orders $p = 1,2,\cdots,P_{max}$ and the corresponding value of the model selection criterion is calculated. Then, the optimal time lag order P can be identified by comparing the scores of the AIC and BIC defined as follows:

$$\text{AIC} = -2\ln(L) + 2k \tag{2.4}$$

$$\text{BIC} = -2\ln(L) + k\ln(T) \tag{2.5}$$

where L is the maximised value of the likelihood function of the model at the value of the parameter estimates, k is the number of parameters in the model, and T is the number of observations. The AIC and BIC scores decrease with an improvement in the log-likelihood while increase with the number of parameters. The lag order with the lowest AIC or BIC score is considered as the best lag order for modelling (Cottrell and Lucchetti, 2016). Since one of the objectives of this study is to reduce the number of parameters in the model, the lowest lag order between AIC and BIC is selected in this study.

The time series variable $\theta_{n,t}$ of predictor link n is considered to Granger-cause the time series variable $\theta_{m,t}$ of target link m if at least one of the lagged values of $\theta_{n,t}$ provides statistically significant information about future values of $\theta_{m,t}$. This can be determined by the F-test on Eq. (2.3) with the null hypothesis $\text{H}_0: \pi_{m,n}^{(1)} = \pi_{m,n}^{(2)} = \cdots = \pi_{m,n}^{(P)} = 0$ and the alternative hypothesis $\text{H}_1: \left(\pi_{m,n}^{(1)} \neq 0\right) \cup \left(\pi_{m,n}^{(2)} \neq 0\right) \cup \cdots \cup \left(\pi_{m,n}^{(P)} \neq 0\right)$. The null hypothesis that $\theta_{n,t}$ does not Granger-cause $\theta_{m,t}$ is rejected if at least one of the elements $\pi_{m,n}^{(p)}$ for $p = 1,2,\cdots,P$ is significantly larger than zero (Bahadori and Liu, 2012). The F-test statistic is computed as follows:

$$F_0 = \frac{\frac{RSS_{restricted} - RSS_{unrestricted}}{k}}{\frac{RSS_{unrestricted}}{(T-2k-1)}} \tag{2.6}$$

where $RSS_{restricted}$ is the residual sum of the squares of the restricted model (i.e., the model with $\pi_{m,n}^{(1)} = \pi_{m,n}^{(2)} = \cdots = \pi_{m,n}^{(P)} = 0$), $RSS_{unrestricted}$ is the residual sum of the squares of the unrestricted model (i.e., the full model in Eq. (2.3)), *k* is the number of restrictions or the number of coefficients being jointly tested, and *T* is the number of observations. For the bivariate GC test, the number of coefficients of the unrestricted model is twice that of the restricted model and, thus, the degrees of freedom in the unrestricted model are equal to $(T - 2k - 1)$. The F_0 value is then compared with the critical value of *F* at the significance level 0.01. The significance level 0.01 is chosen, as it is more conservative in reducing Type I error (false positive finding) compared to the more commonly used 0.05 significance level. In several studies related to Granger causality test such as Vanco (2012), Yu et al. (2015), Chu et al. (2016), Hasan and Kim (2016) also considered significance level 0.01. If the F_0 value is higher than the critical value obtained from the *F* table at the significance level 0.01, the null hypothesis is rejected and the time series of the tested predictor link *Granger-causes* the time series of the target link. We call *GC links* of the target link the predictor links that Granger-cause a given target link.

2.2.2 Granger-causal strength for predictor ranking

The causal relation between the two variables can be further quantified by computing the 'Granger-causal (GC) strength' (Geweke, 1984; Bressler and Seth, 2011) to measure the strength of the causal relation between a predictor variable and the response variable: the higher the GC-strength, the higher the relevance of the response variable.

The GC-strength is defined as the logarithm of the ratio of the variability of the residual of the restricted model $(\varepsilon_{restricted})$ to the variability of the residual of the unrestricted model $(\varepsilon_{unrestricted})$ (Geweke, 1984; Bressler and Seth, 2011):

$$S_G = \ln \frac{Variance\ of(\varepsilon_{restricted})}{Variance\ of(\varepsilon_{unrestricted})} \tag{2.7}$$

where S_G denotes the GC-strength of a predictor variable on the target variable. Eq. (2.7) can also be expressed in terms of the mean squared error (MSE) and residual sum of squares (RSS) as shown in Eq. (2.8) and Eq. (2.9) below, respectively:

$$S_G = \ln \frac{MSE_{restricted}}{MSE_{unrestricted}} \tag{2.8}$$

$$S_G = \ln \frac{\frac{RSS_{restricted}}{(T-k-1)}}{\frac{RSS_{unrestricted}}{(T-2k-1)}} \tag{2.9}$$

where $(T-k-1)$ are the degrees of freedom of the restricted model. The GC-strength cannot be negative (Bressler and Seth, 2011) and can be measured by using F statistics (Geweke, 1982; Thalassinos et al., 2015).

For a given target link, we first identify its GC links by conducting the F-test using Eq. (2.6) and then measure the magnitude of causal effect of each GC link using Eq. (2.9). As the importance of the predictors increases with their GC-strength, we then select systematically the *K* most important predictors for a particular target link. In this study, we propose the use of a *percentile* to set a cut-off value of GC-strength and select the associated predictors. For instance, if we use the 90th percentile GC-strength as a threshold, we select predictor links whose GC-strength is greater than the threshold as input variables, meaning that only the top 10% of GC links will be included in the prediction model of the target link. Similarly, if we use the 80th percentile GC-strength as a cut-off, we will have the top 20% of GC links included in the model. The higher the percentile number is, the more parsimonious the model is, but the less accurate the prediction result would be as a lesser number of parameters will be used in prediction. Finding a proper level of percentile cut-off is important as it requires a trade-off between model complexity and model accuracy. In this study, we test different percentile levels and provide recommendations for using a variable selection method based on GC-strength. The test results and recommendations are provided in the case study below.

2.2.3 Granger causality method evaluation

As the proposed GC-based method is an input variable selection method for a short-term traffic prediction model, its performance should be evaluated based in terms of its contribution to improving the prediction accuracy of a given traffic prediction model. The evaluation procedure consists of the following steps:

1. For a given target link, determine the GC links and GC-strength cut-off to select a set of predictor links.
2. Build a short-term traffic prediction model using the model type of choice (e.g., time-series regressions, neural networks).
3. Measure the prediction accuracy of the short-term prediction model and evaluate the effectiveness of the most important GC predictors in improving the model performance.

In short-term traffic prediction, the forecasting horizon is typically less than an hour (Smith et al., 2002; Pascale and Nicoli, 2011). In this study, the 5-minute time horizon is used and the short-term traffic prediction is implemented using two types of prediction models: time-series regression and multi-layer feed forward neural network.

For the first type of prediction model, we consider a simple time series model based on multiple linear regression method. The traffic state of the target link is taken as the response variable and past traffic states of the most important GC links as well as the past traffic state of the response variable itself are taken as explanatory variables. The time step is set to 5 minutes and the time lag of the set of predictors is computed by using Eq. (2.5). The parameters of the regression are estimated by using 80% of the data (training set) and the prediction performance is computed by using the remaining 20% of the data (testing set).

For the second type of prediction model, we consider a fully-connected feedforward artificial neural network with back-propagation learning algorithm, which is known as multi-layer perceptron (Bishop, 1995). A simple multi-layer perceptron includes an input layer, one or more hidden layers, and an output layer. Each layer contains one or more neurons, and the neurons are connected by directed edges which are known as synapses. The synapses are associated with a weight representing the strength of the connection between two neurons. In this study, the input layer contains neurons representing traffic states of the most important GC links, one hidden layer connects input and output, and the output layer has one neuron representing the traffic state of the given target link. It should be noted that the number of neurons in the input layer varies with the target link, as a different number of important GC links is obtained for different target links, and hence the total number of neurons in the hidden layers also varies. For simplicity of

the prediction model, the total number of neurons in the hidden layer is selected by trial and error. The package 'neuralnet' in R programming language (Günther and Fritsch, 2010) is used to build the neural network. Similar to the linear regression, the dataset is partitioned into an 80% training and a 20% testing dataset. It should be noted that the data are normalised based on the min-max method and are scaled in the interval of [-1, 1] to facilitate the model implementation.

Once the models are estimated, their prediction accuracy for each target link m is measured by using the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE):

$$RMSE_m = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\theta_{m,t} - \hat{\theta}_{m,t}\right)^2} \tag{2.10}$$

$$MAE_m = \frac{1}{T}\sum_{t=1}^{T}\left|\left(\theta_{m,t} - \hat{\theta}_{m,t}\right)\right| \tag{2.11}$$

where T is the number of observations, $\theta_{m,t}$ is the observed value of a traffic state of target link m at time t, and $\hat{\theta}_{m,t}$ is the associated model predicted value.

2.3 Case study

2.3.1 Road network

The urban road network in Brisbane (Australia) is selected as the site for a case-study. The network consists of freeways, highways, arterials and minor road links that traverse the Brisbane central business district (CBD) and surrounding areas. Among all the road links in the study site, 522 road links have available traffic observations collected from loop detectors. However, due to significant amount of missing data and invalid traffic measurements of malfunctioned detectors, 43 links among 522 road links have been excluded from this analysis. Therefore, traffic measurements of 479 road links are considered as variables in this study. The bivariate method of spatial variable selection was implemented for six road links that were considered as target links while all 479 links were analysed as potential predictor links. The six selected target links differ in terms of location on the network and road types: target links 2, 5, and 6 are freeways whereas target links 1, 3, and 4 are arterial roads. Figure 2.1 shows the road network map with the locations of the 479 road links including the selected six target links.

2.3.2 Traffic state data

Traffic flow measurements for the 522 road links on the selected Brisbane network were obtained from the Queensland Department of Transport and Main Road (DTMR) through the Public Traffic Data System (PTDS). The data contained 3-minute traffic volume and speed measurements from the loop detectors of 522 road links for the duration of 7 months from May 1, 2016 to November 29, 2016. 3 minute traffic volume and speed data are converted to 5 minute interval traffic flow and speed data. Since the whole day period (24 hours of each day) is considered for the analysis in this study, the total number of observations of the continuous time series flow data of each road link is 61,344. It should be noted that at times the detectors were not working properly, and hence missing values or invalid values in traffic flow measurements appeared in the data. Among 522 road links, the links which have more than 40% missing data for the duration of 7 months are omitted to investigate statistical causal relations between link flow measurements. Therefore, flow measurements of 479 road links are selected as the variables in this study. On average, 7% missing data are observed in the selected 479 road links which required to be imputed. To fill in the data used for analysis, this study used multiple imputation (Sterne et al., 2009). In multiple imputation, the fully conditional specification method (Lee and Carlin, 2010) is adopted in which a specific univariate model (i.e. linear regression) for each variable with missing values is used to impute the missing values by taking all other available variables as the predictors. This method iteratively imputes each variable with missing values, then uses the imputed values in the imputation of missing values of other variable and continues until reaches the maximum iteration. Multiples imputation was implemented using software package 'SPSS' (SPSS Inc., 2008) where the maximum number of iteration was taken as 10 and the average value of five imputations was considered as the imputed value.

As mentioned in the methodological section, the GC test requires that time series variables are stationary. However, traffic flow is a non-stationary process (Pascale and Nicoli, 2011; Li et al., 2015), and hence we applied *de-trending* to remove non-stationarity (e.g., time-of-day and day-of-week periodicity) from the time series, where for each observation the average traffic flow of the corresponding day of the week and time of the day is subtracted. This gives a time series of residuals from the mean traffic flow of the same time-of-day and day-of-week, which makes the time series data stationary. The stationarity of the time series was assessed via the Augmented

Dickey Fuller unit root test, which rejected the null hypothesis at the 95% confidence level and hence verified that the data fulfil the condition of a stationary process and every variable is within unit root of 1. Also, the Durbin Watson test (Durbin and Watson, 1971) at the 95% confidence level showed that the model residuals were not auto-correlated.

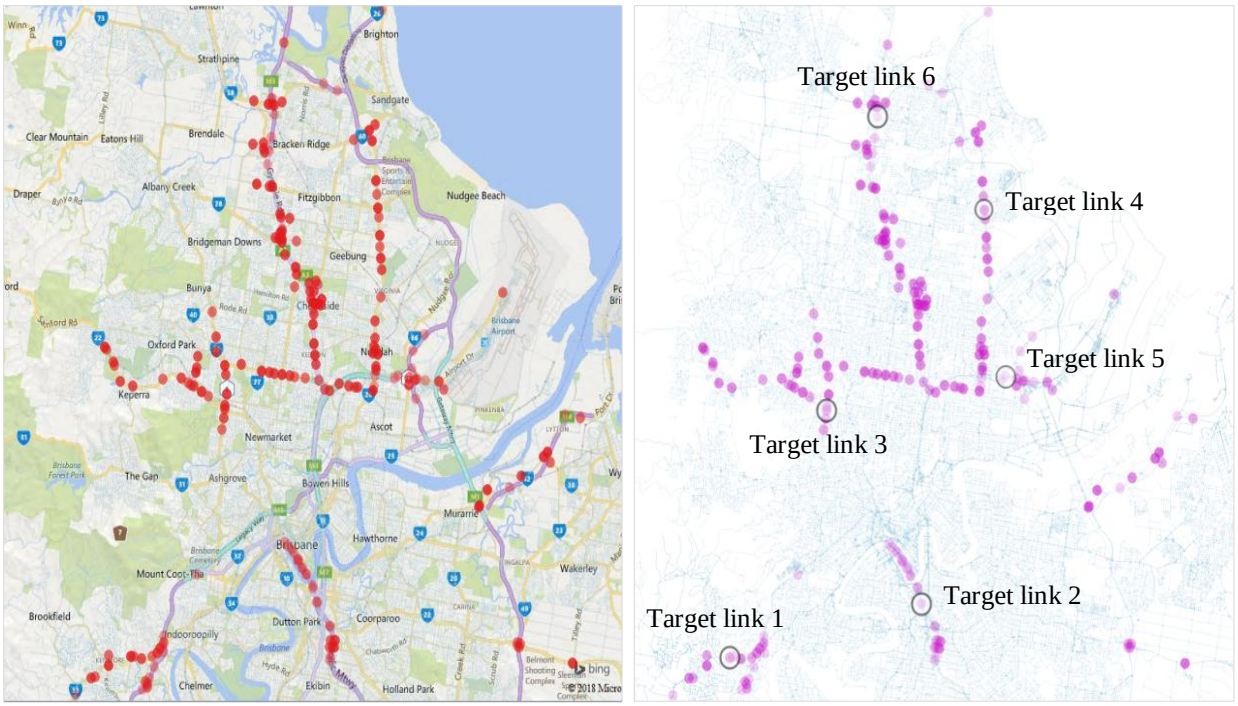


Figure 2.1: (a) Road network and links with traffic measurements, (b) Target road link locations.

2.3.3 Variable selection scenarios

The proposed variable selection method was implemented to find the GC links and measure their GC strength to rank the most important links within a GC-strength percentile. With the aim of evaluating the effectiveness of the proposed method in short-term traffic prediction applications, we built prediction models while considering different variable selection methods and compared the resulting prediction performances. For each of the six target links, a total of five variable selection scenarios were considered according to the input variables of the prediction models:

1. The lagged values of the target link itself.
2. The lagged values of the target link and the white noise error terms.
3. The nearest upstream and downstream links of the target link.

4. The set of links in the neighbourhood of the target link, with a number of predictors equal to the ones in the selected GC-strength percentile.
5. The set of GC links with a number of predictors equal to the selected GC-strength percentile.

The five scenarios reflect different approaches to consider spatial dependence, with Scenario 5 corresponding to the implementation of the proposed GC-based variable selection method. Scenario 1 is the simplest model that does not consider spatial dependence but only the history of the traffic flow variables of the target link. Scenario 2 adds the history of the error terms to the previous scenario, as the future value of a target variable depends on its own lagged values, captured by autoregressive (AR) terms, and the prediction error is a linear combination of current and past prediction errors, captured by moving average (MA) terms (Ahmed and Cook, 1979). As the orders of AR and MA depend on the time series of traffic flow of each target link, different orders of AR and MA are observed for different target links. Therefore, the orders of AR and MA are not taken as same fixed numbers for all target links. Scenario 3 considers spatial dependence being limited to the road links located in close proximity to the target link, while Scenario 4 enlarges the amount of road links considered to the same number of predictors selected in Scenario 5. In particular, the comparison of Scenarios 4 and 5 will allow us to evaluate the importance of integrating the spatial dependence structure in selecting input variables for traffic predictions when the same number of predictors is used.

2.4 Results

We present the results of the implementation of the GC-based method: the GC links are first selected and then ranked according to GC-strength, so that the set of the most relevant predictors can be used as the input for any prediction model. We also discuss the temporal and spatial characteristics of the GC links and compare results across the five aforementioned scenarios.

2.4.1 GC-based method implementation

The implementation of the spatial variable selection method identified GC links for the six target links by considering initially all the 479 road links in the road network as potential predictor

links. Then, the GC-strength of the predictors was computed and Figure 2.2 shows its frequency distribution for all the six target links: the distribution is heavily right-skewed, with mean equal to 0.36×10^{-2} and most values between 0 and 0.25×10^{-2}. Figure 2.2 presents also the cumulative distribution and shows the 50th, 75th, and 90th percentile values for the selection of the number of links as described in Scenario 5.

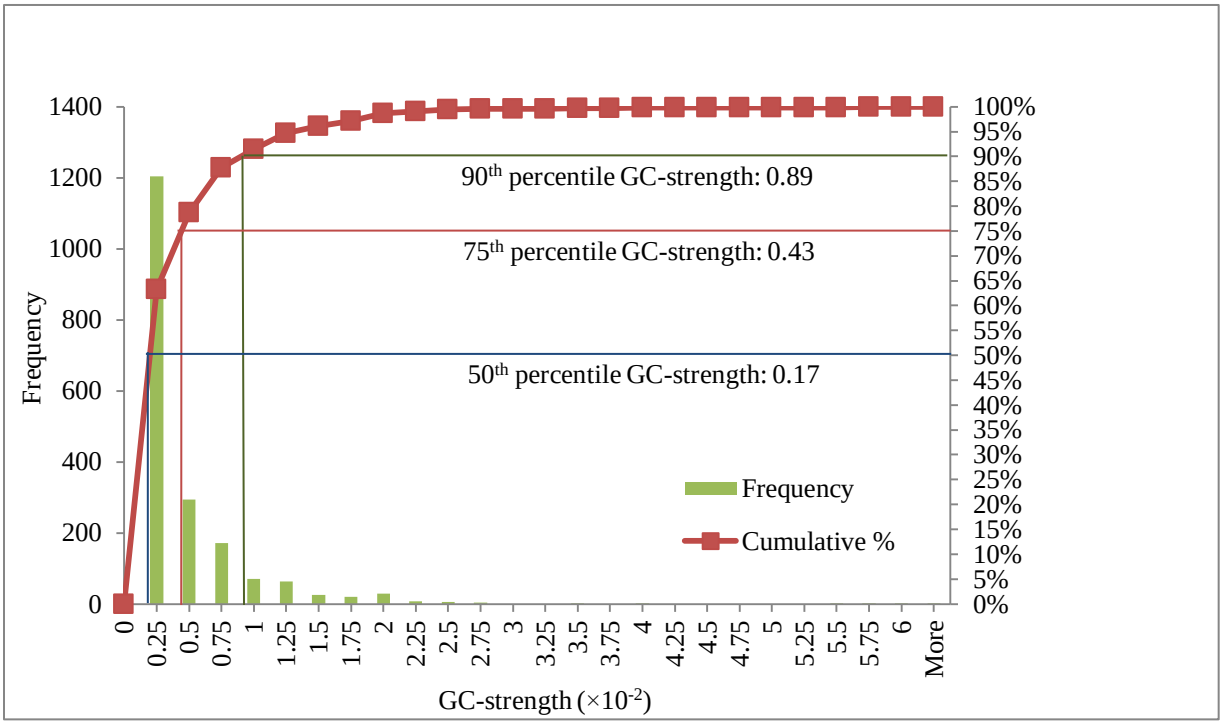


Figure 2.2: Frequency distribution and cumulative (%) frequency distribution of GC-strength for all target links.

Considering the three different percentile values, we determined the GC-strength threshold value for our variable selection method. For each target link, a time-series regression-based traffic prediction model was built by using predictor links within the three percentile values. For instance, the prediction model with predictors within the 75th percentile threshold included 25% of all GC links whose GC-strength was greater than or equal to 0.43×10^{-2}. The model performance was measured by comparing prediction accuracy and model simplicity via the BIC. For the linear regression-based prediction model, $\mathrm{BIC} = T\ln(MSE) + k\ln(T)$, where $MSE = \frac{1}{T}\sum_{t=1}^{T}\left(\theta_{m,t} - \hat{\theta}_{m,t}\right)^2$ for target link m.

Figure 2.3 shows the BIC results for the six target links for the 50th, 75th, and 90th percentile threshold levels. The BIC values are consistently the lowest for the 90th percentile threshold, indicating that the GC-strength threshold of the 90th percentile performs the best when considering both prediction accuracy and model simplicity. It should be noted that the same consistency is not observed when comparing the 50th and 75th percentiles, as the latter does not perform the best for all links. Based on these results, we selected the 90th percentile as the GC-strength threshold for the proposed variable selection method, and we considered all GC links with GC-strength, greater than or equal to 0.89×10^{-2} as input variables for the prediction model (either multiple linear regression or neural network) of each target link.

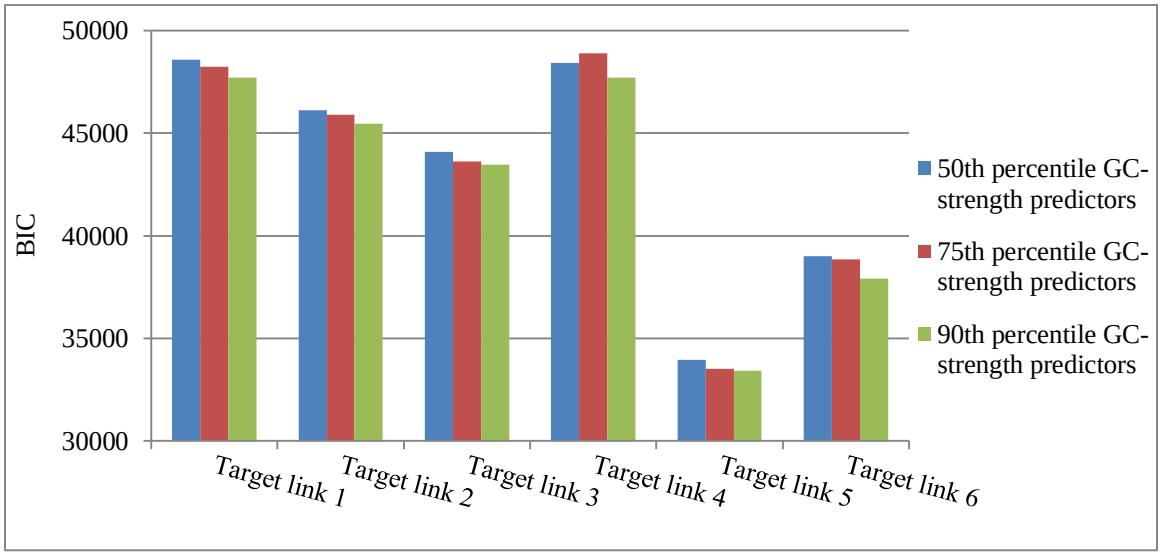
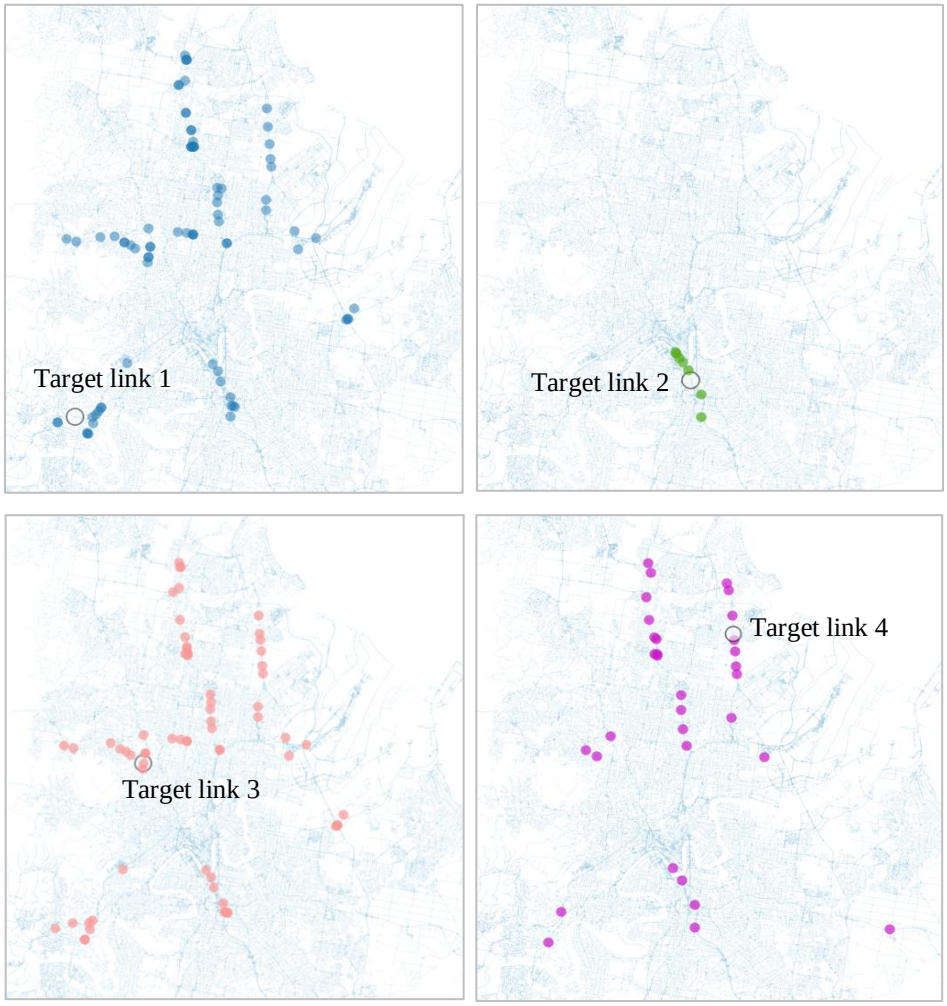


Figure 2.3: Comparison of BIC values for GC-strength threshold values.

2.4.2 Dimensionality reduction

Having selected the 90th percentile as the GC-strength threshold for the proposed GC-based method, the analysis of the number and location of the GC links for each target link offers insight into the amount and spatial distribution resulting from the dimensionality reduction. Figure 2.4 presents the number and the location of the GC links within the 90th percentile for the six target links, and the immediate consideration is that the dimensionality reduction is substantial when considering that about 7% of total road links were on average selected as GC links for a target link. Moreover, a difference exists in the amount of predictors between links 1, 3 and 4 (with a

large number of GC links per target) and links 2, 5 and 6 (with a small number of GC links per target). Notably, the latter group is made of freeway links, indicating that the GC-based method selects a lower number for this particular type of road links.



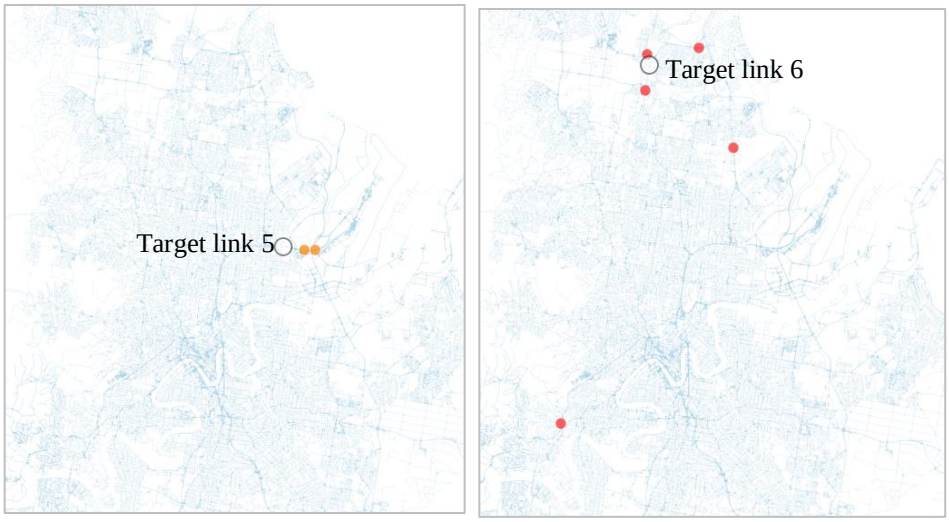


Figure 2.4: GC links for each target link within the 90th percentile GC-strength threshold.

When looking at the location of the GC links for the target links, differences can be observed for not only the two aforementioned groups, but also within the freeway link group. Clearly, the proposed method based on the 90th percentile GC-strength threshold produced different spatial distributions of the relevant predictors for each of the target link.

The first important observation is that the predictor links are often located far from the target link. The proposed approach selects a set of predictor links whose time series provide statistically significant information about future time series of the target link regardless of the distance between the target link and the predictor links. Traffic states of the selected predictor links which are located far from the target link may not have direct impact on the traffic states of the target link; however, the traffic states of these predictor links can provide indirect information about the future traffic states of the target link. A distant predictor links can be statistically correlated to the target link when they possess similar traffic states in response to a common network-wise cause such as demand level, weather, holidays etc. Changes in the traffic states of the predictor link can provide an indication of similar changes in the traffic states of the target link. Thus, a predictor link can provide information that helps the estimation of the traffic state of a target link even in the cases that the link is not physically connected or not even in close proximity to the target. Therefore, the proposed GC-based approach extends the spatial boundaries of traffic prediction to be network-wide, rather than limiting them to traditional

approaches of considering only neighbouring links, if not even physically connected ones like the upstream and downstream links.

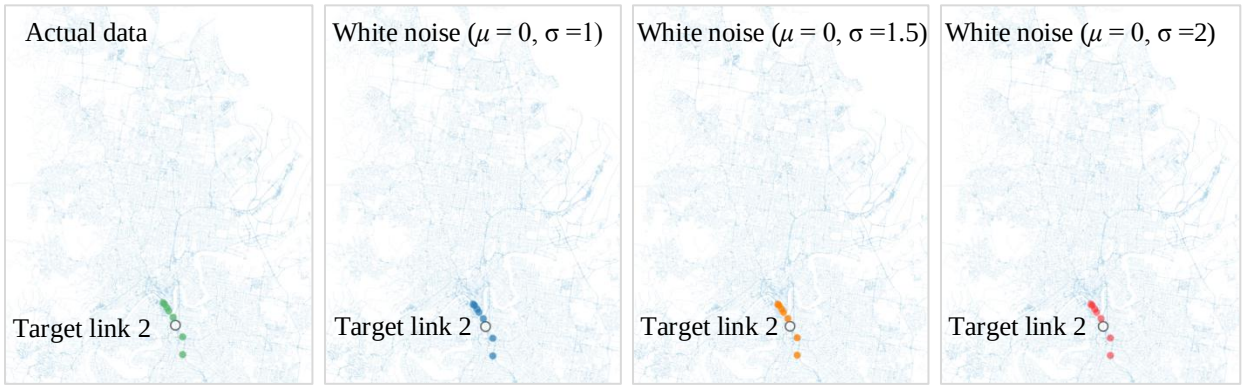
A second important observation is that, even though the spatial distribution of the GC links covers the entire network, our findings show that upstream and downstream links of target link 2 and target links 5-6 are selected as important predictor links by the proposed spatial variable selection method. As aforementioned, this group of links contains freeway target links and the results suggest that the upstream and downstream links of the freeway target link have higher GC-strength and hence these links will have a stronger causal on the target link. In other words, the more traditional approach to select neighbouring links might apply to a specific type of links rather than the entire network.

2.4.3 Characteristics of GC-based method

2.4.3.1 Robustness in selection of the significant predictor links

The robustness of selecting the most significant predictors by the proposed method in case of error in traffic states measurements is also investigated. This study determines whether a minor change in traffic states data could change the selection of the most significant predictors for a target link. For this purpose, random error in traffic states time series data i.e. white noise (with mean, $\mu = 0$ and standard deviation, $\sigma = 1, 1.5, 2$) is added to flow time series of all of the road links. Flow time series data used in this study are de-trended data (which includes positive and negative values). We applied de-trending to remove non-stationarity (e.g., time-of-day and day-of-week periodicity) from the time series data, where for each observation the average traffic flow of the corresponding day of the week and time of the day is subtracted. The typical mean value of the flow time series de-trended data of a road link is approximately -5 (Target link 2 as an example). Therefore, adding white noise (with $\mu = 0$ and $\sigma = 1, 1.5, 2$) to the flow time series data would have minor changes in the traffic states data. The proposed method (90^{th} percentile GC-Strength based variable selection) is then applied to the traffic states data including the white noise. The most significant predictors for each target link selected by the proposed method based on the traffic states data including white noise and excluding white noise (i.e. actual data) is compared. In both cases, the most significant predictors selected for each target link by the

proposed method are same which indicates the proposed method is robust in context of errors in measurements. Figure 2.5 illustrates the locations of the selected the most significant predictors for a target link (Target link 2 as an example) based on the actual traffic states data and the traffic states data including white noise of different standard deviations (σ =1, σ =1.5 and σ = 2). It is observed from the figures that the selected significant predictors for the target link is same regardless of the white noise added to the traffic state data.



(a) (b) (c) (d)

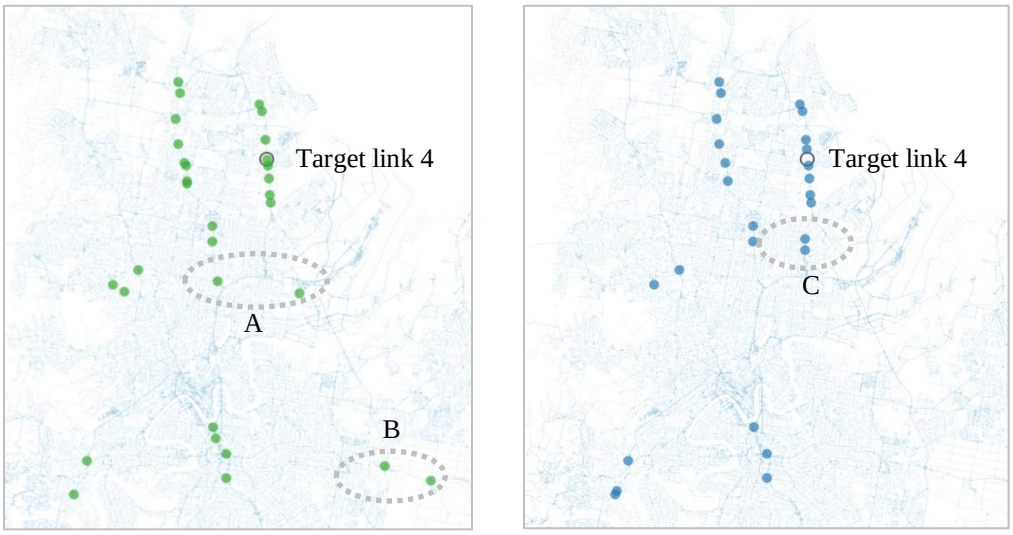
Figure 2.5: Significant predictor links for a target link (Target link 2 as an example) within the 90th percentile GC-strength threshold of flow data when random error in measurements (white noise) is considered.

2.4.3.2 Evolution of the selected significant predictor links over time

This study also explores whether the significant predictor links for a target link identified by the proposed 90th percentile GC-Strength based variable selection method change over time. The seven months traffic states data is divided into two segments where each set contains three months of traffic states data and the proposed method is applied in two data sets: i) traffic states data for the month of May to July and ii) traffic states data for the month of August to October.

After applying 90th percentile GC strength based variable selection method on two data sets separately, a significant number of common links are obtained in both data cases. Figure 2.6 represents the locations of the selected most significant predictor links for a target link (Target link 4 as an example) in two data cases. It can be observed that the most of the predictor links

identified in two cases are same. However, the predictor links located in region A and B in Figure 2.6(a) cannot be observed in Figure 2.6(b) and similarly, the predictor links located in region C in Figure 2.6(b) cannot be seen in Figure 2.6(a). Therefore, some significant predictors identified by the proposed method based on first three months data (May to July) are not noted by the proposed method based on next three months data (August to October). As the proposed method is a data-driven method, the selection of the significant links depends on the data set used. Since these two sets of data have traffic state data of different months, the selection of the relevant links by the proposed method varies slightly.



a) Data: May to July b) Data: August to October

Figure 2.6: Significant predictor links for a target link (Target link 4 as an example) within the 90th percentile GC-strength threshold based on three months data.

2.4.3.3 Variation of the selection of significant links in peak and off-peak periods

The proposed 90th percentile GC-strength based spatial variable selection method is implemented in different time of the day to analyse whether the selection of the significant links for a target links varies over time. For this purpose, traffic states data of whole day period is divided into four parts based on the time of the day which are morning peak period (7 a.m. - 11 a.m.), daytime off-peak period (11 a.m. - 4 p.m.), afternoon peak period (4 p.m. - 8 p.m.) and night time off-peak period (8 p.m. - 7 a.m.). Figure 2.7 illustrates the average traffic flow of each six target link during peak and non-peak periods of seven months (May 2016 - November 2016).

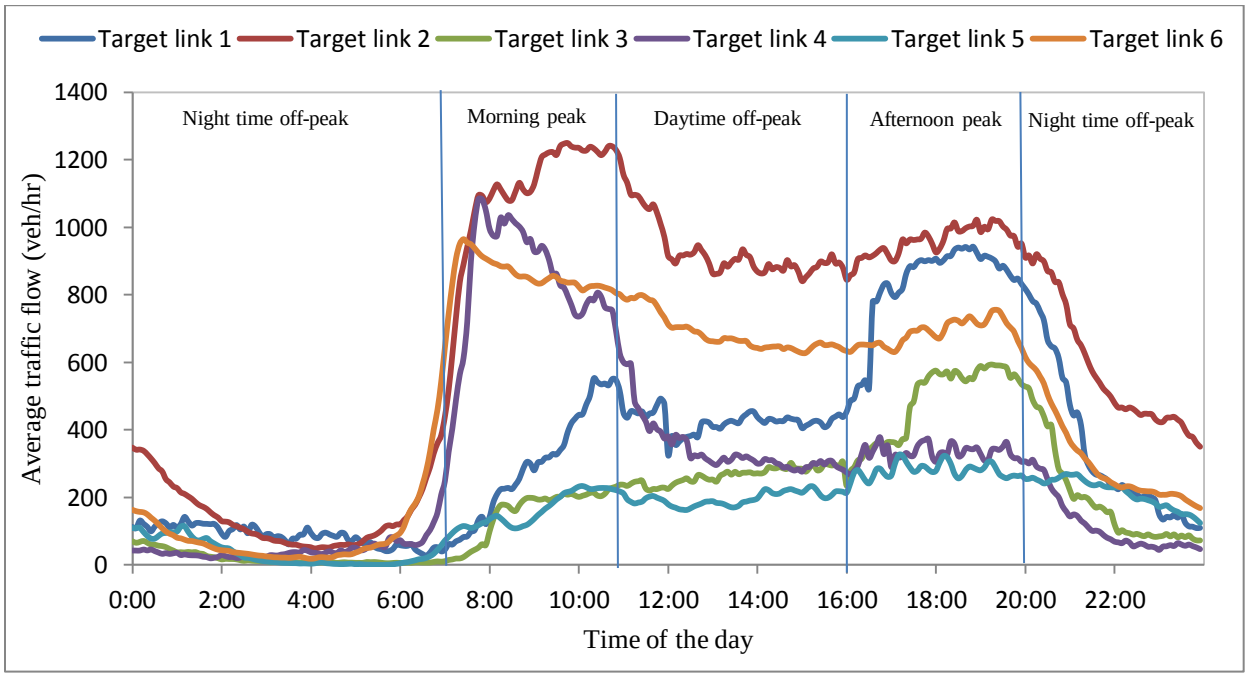
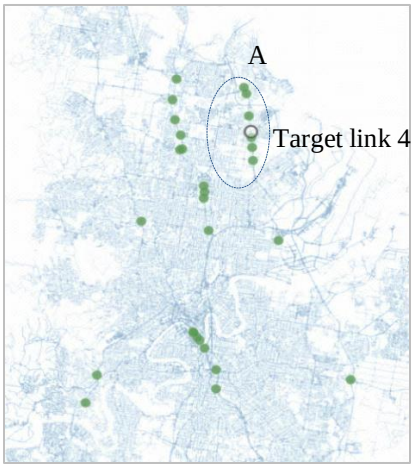
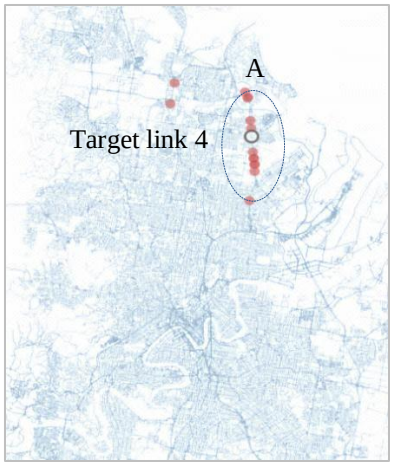


Figure 2.7: Selection of peak period and off-peak period of six target links by using average traffic flow.

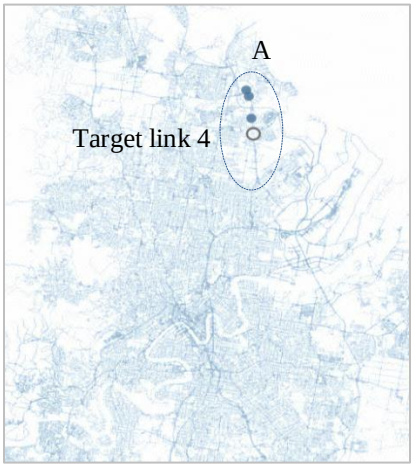
The analysis results demonstrate that the spatial connectivity is dependent on time of the day. Figure 2.8 depicts the location of the selected significant predictor links for a target link (Target link 4 as an example) by the proposed method based on the traffic states of the whole day period, morning peak period, daytime off-peak period, afternoon peak period and night time off-peak period. As the proposed method is a data-driven method, the selection of the significant predictor links depends on the data set used. Since four time periods (morning peak period, daytime off-peak period, afternoon peak period and night time off-peak period) have different set of traffic state data, the selection of the significant predictor links by the proposed method varies. However, some neighbour links (region A in Figure 2.8) of the target link are selected as the significant predictor links in these four time periods. This indicates that some neighbour links possess high GC-Strength at all time periods of a day. Also, the selected significant links in the four time periods are also selected by the proposed method when the traffic states of whole day period is considered because whole day traffic state dataset includes peak and off-peak period traffic state data.



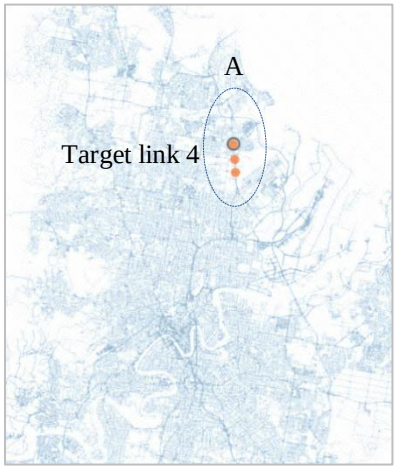
(a) Morning peak period



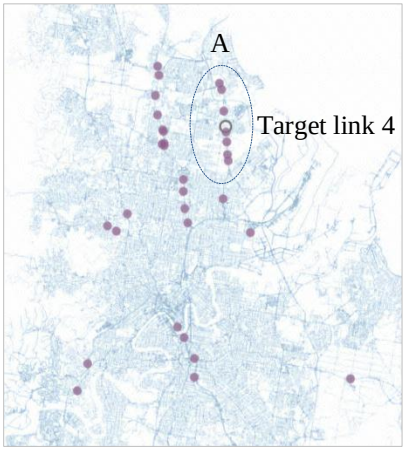
(b) Daytime off-peak period



(c) Afternoon peak period



(d) Night time off-peak period



(e) Whole day period

Figure 2.8: Significant predictor links for a target link (Target link 4 as an example) within the 90th percentile GC-strength threshold during whole day, peak periods and off-peak periods.

2.4.3.4 Selection of the significant predictor links by different traffic parameters

This study evaluates the variation of identifying the most significant predictors for each of the six target links by the proposed method when traffic flow and speed are considered separately as the traffic state. The results shows that although having some common links, a number of significant predictor links selected using traffic flow data as the traffic states are different from the significant predictor links selected based on speed data as the traffic state. This indicates the selection of the significant predictors depends on the traffic parameters (flow or speed) that are used as the traffic state. These two traffic parameters have different ranges of values. Also, in a fundamental diagram of speed-flow, a single value of flow can be obtained in congested and uncongested traffic regime and it represents two different speed values. This is one of the key factors of selecting a number of different significant predictors for a target link when two different traffic parameters are used.

The proposed method of this study focuses on a single traffic parameter to estimate the traffic states of the road links in the network. For instance, traffic parameter namely traffic flow is used as the traffic states of the road links when traffic flow of the target link are to be predicted. Similarly, traffic parameter namely speed is used as the traffic states of the road links when speed of the target link are to be predicted. Therefore, the selection of the significant predictor links for a target link by the proposed method depends on the traffic parameter used. However, future research will consider using multiple parameters together in the proposed method to identify a more generalised set of significant predictor links for the target link.

Figure 2.9 shows the locations of the most significant predictors for the six target links selected by 90^{th} percentile GC strength based variable selection method based on speed data. Figure 2.10 represents the number of the most significant predictors that can be obtained when traffic flow or speed is considered as the traffic state. They are the most common significant predictors for each target link by the proposed method regardless of the traffic parameter data (traffic flow and speed) used for analysis. It can be observed from Figure 2.9 and Figure 2.4 that most of the common significant predictors are located nearer to the target link.

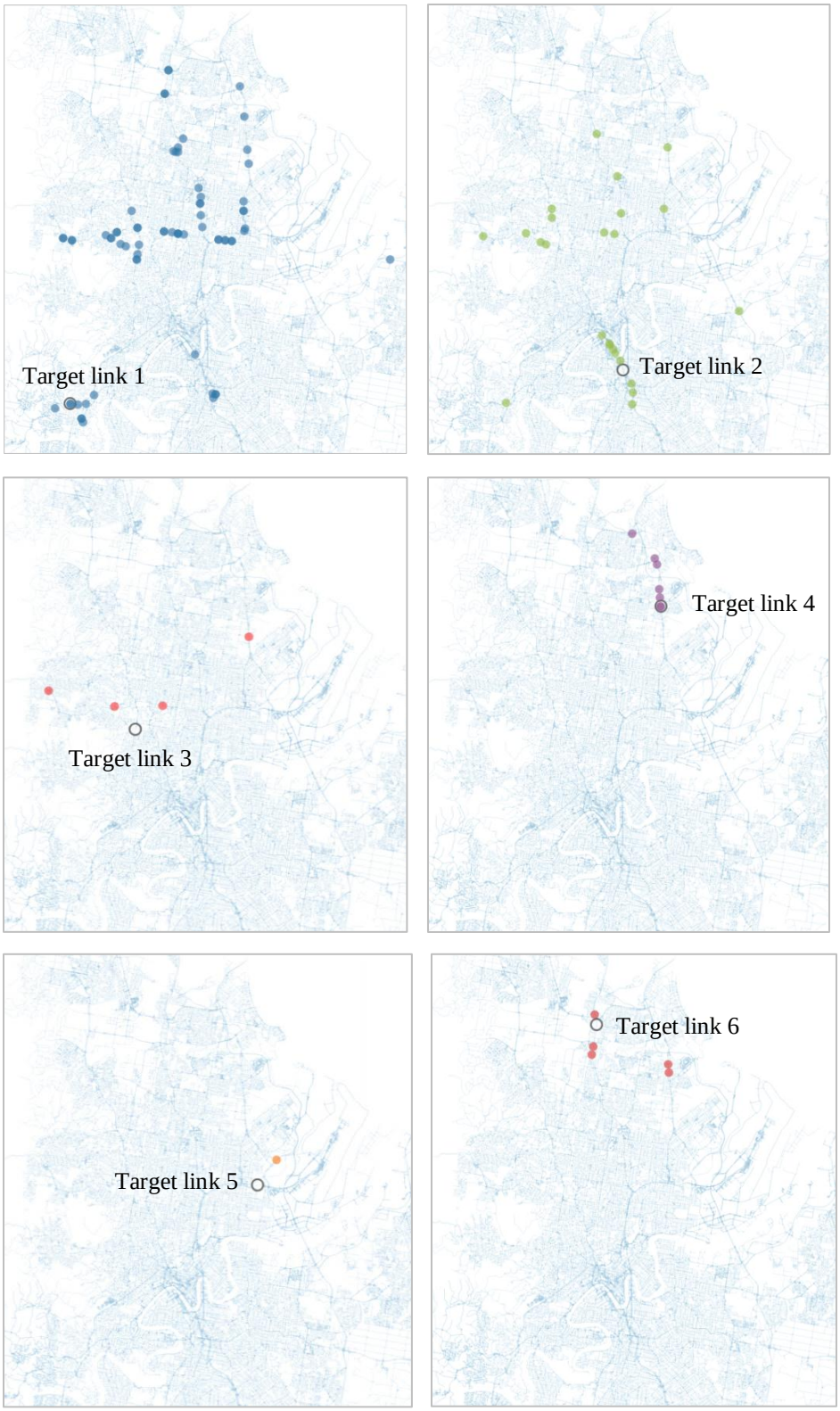


Figure 2.9: Significant predictor links for each target link within the 90th percentile GC-strength threshold when speed data are considered.

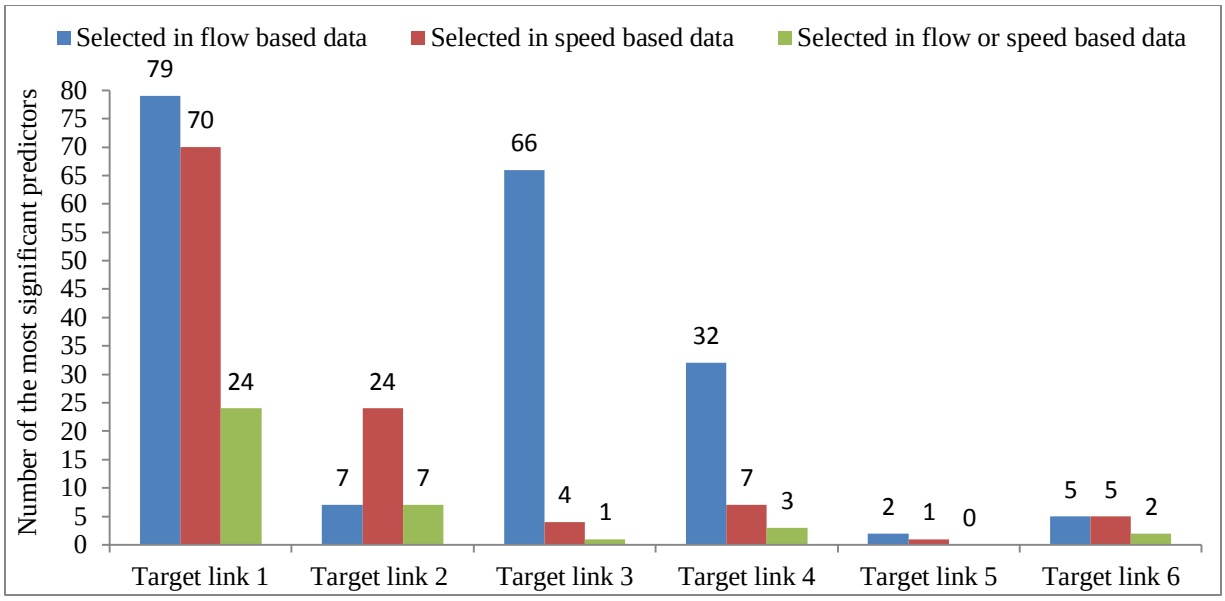


Figure 2.10: Number of significant predictor links for six target link within the 90th percentile GC-strength threshold of flow data and speed data.

2.4.4 Temporal characteristics of GC links

2.4.4.1 Time-of-day trends

As the Granger-causal relation between two road links exists if the time series of one link has causal effects on the time series of the other, the patterns of these time series are closely related and their difference is approximately constant over time (Ahmad and Hurnhirun 1995; Demirbas 1999). Accordingly, it is expected that links with Granger-causal relations would share similar time series patterns that can be visually identified.

We examined the time series plots of traffic flow for the GC links for a selected target link to understand what time series characteristics are captured by the GC test and whether there is any visually identifiable difference in time series patterns between GC links and non-GC links. We illustrate our findings by showing the results for target link 4 for this analysis and visualising three GC links (Link ID = 38, 288, 294) and three non-GC links (Link ID = 144, 324, 351), whose locations are presented in Figure 2.11. It should be noted that both the GC links (i.e., relevant predictors of the traffic state of link 4) and the non-GC links are not in close proximity

of the target link, and also that GC and non-GC links are in proximity of each other, suggesting the need of a closer look at their characteristics.

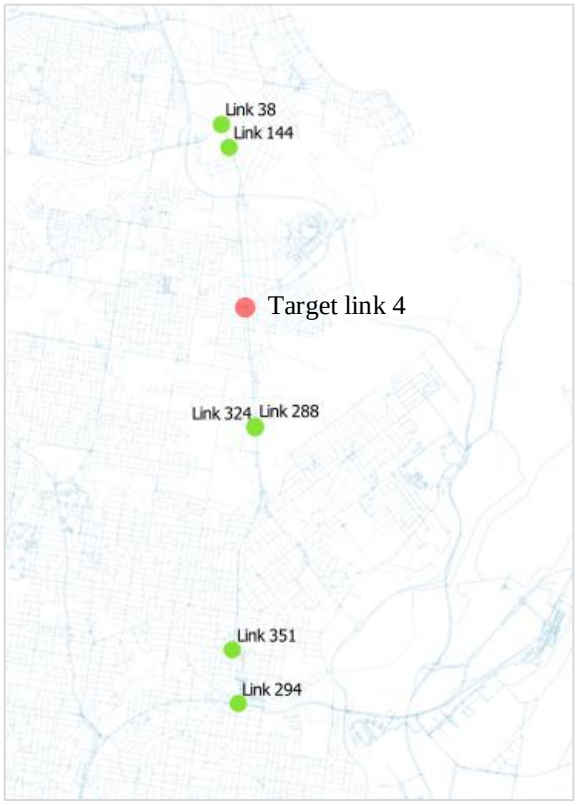
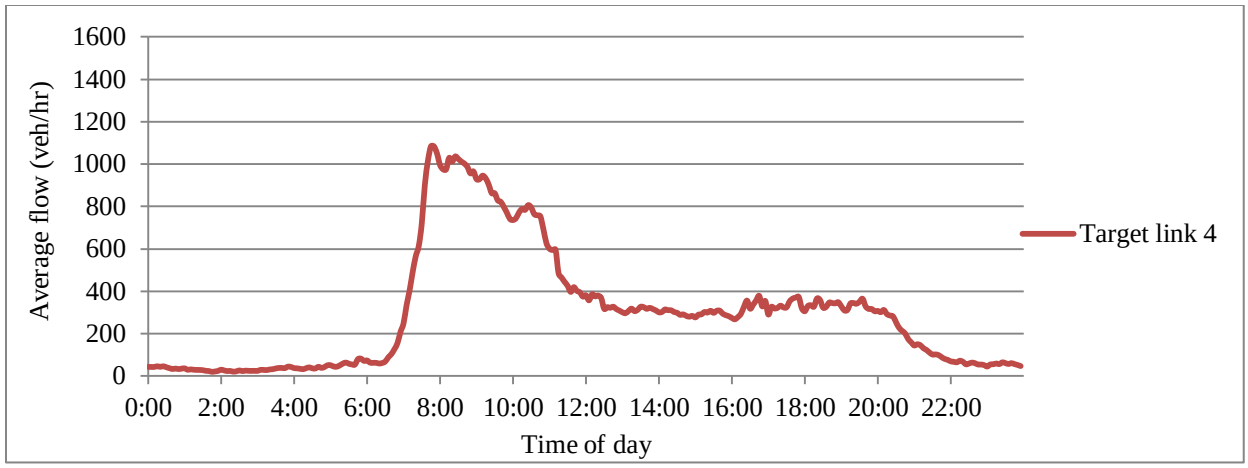
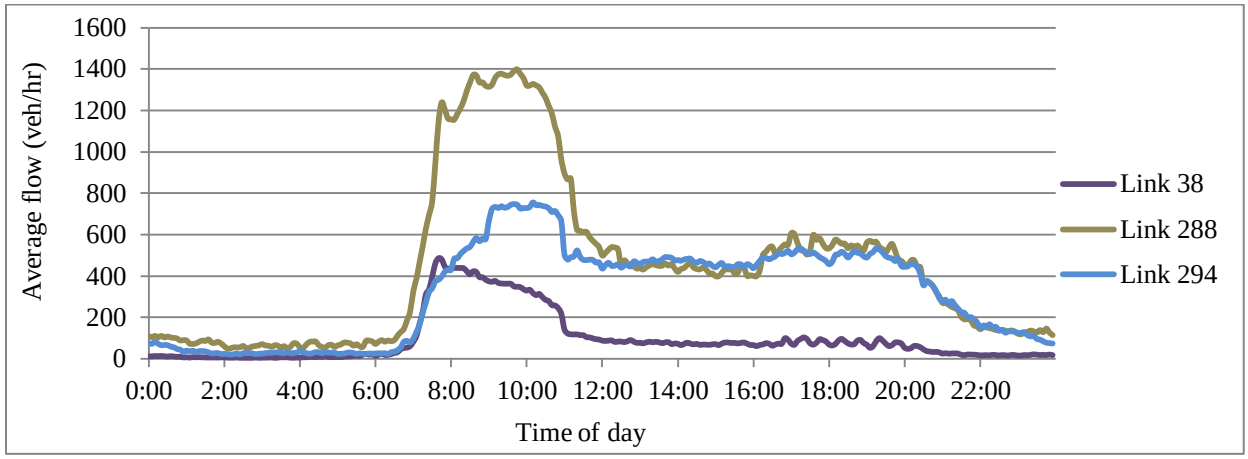


Figure 2.11: Location of the target link and the set of predictor links.

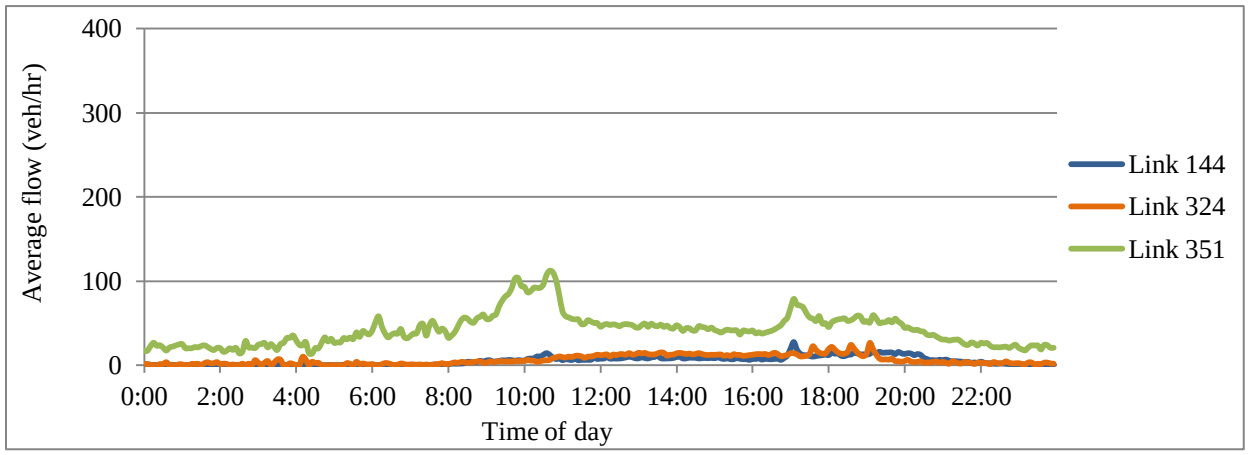
Figure 2.12 illustrates the average traffic flow patterns of the target link, the GC links, and the non-GC link. Given the 5-minute interval measurements from the dataset, traffic flow observations are the result of the average across multiple days for each 5-minute interval of the day. The patterns show visually the temporal characteristics of Granger-causal relations: the comparison of patterns in Figures 2.12(a) and 2.12(b) shows that the GC links present similarities to the target link, in particular in terms of timing and duration of peak and non-peak hours; the comparison of patterns in Figures 2.12(a) and 2.12(c) illustrates instead that there exist notable differences between the target link and the non-GC links. It should be noted that the presentation of the results for target link 4 applies to all relations between target links and GC or non-GC links.



(a) Target link



(b) GC links of the target link



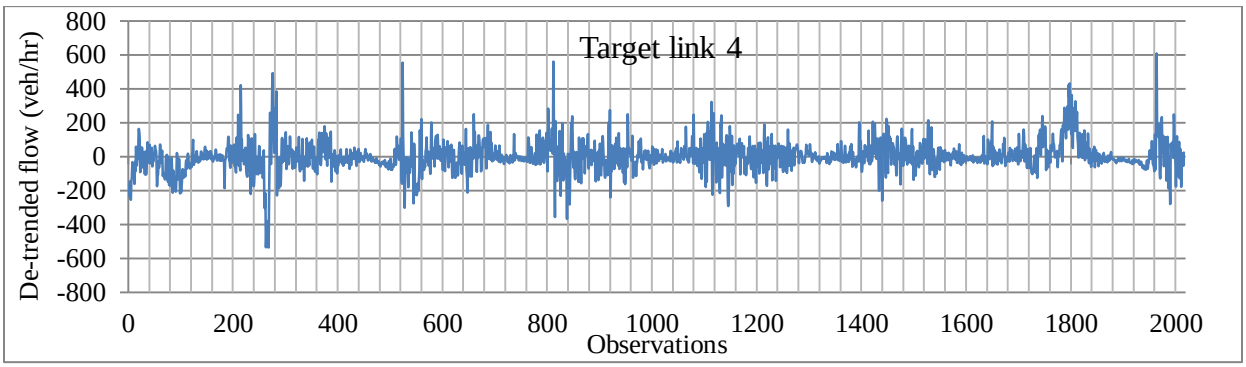
(c) Non-GC links of the target link

Figure 2.12: Average traffic flow patterns of a target link, GC links and non-GC links.

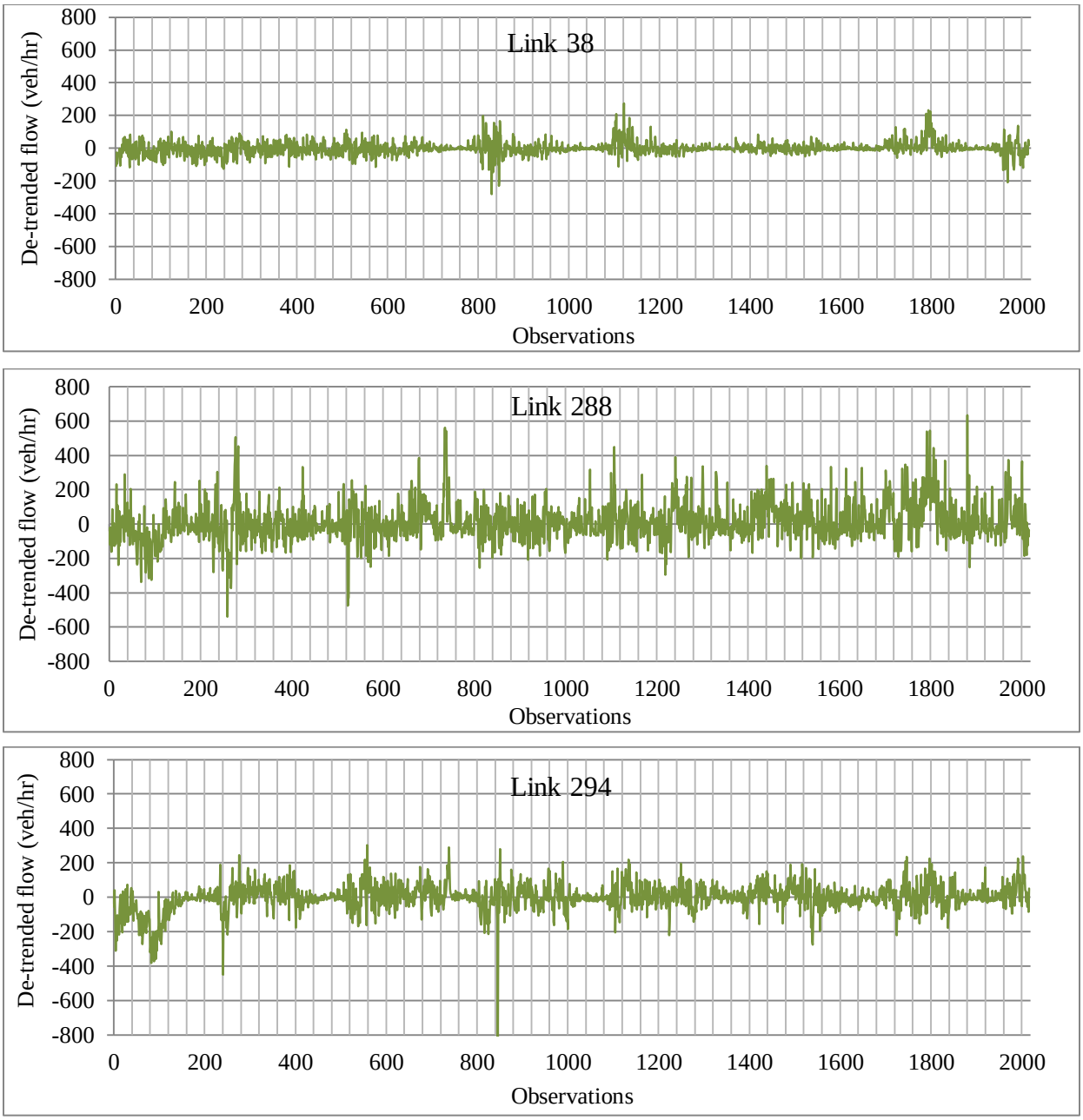
2.4.4.2 Day-to-day fluctuations

The GC test required for the time series to be stationary and the data were de-trended to implement the variable selection method. Looking at the de-trended data, namely the residuals after the removal of historical time-of-day mean trend, allows to investigate the differences between GC links and non-GC links of the target links in terms of day-to-day fluctuations around the long-term mean.

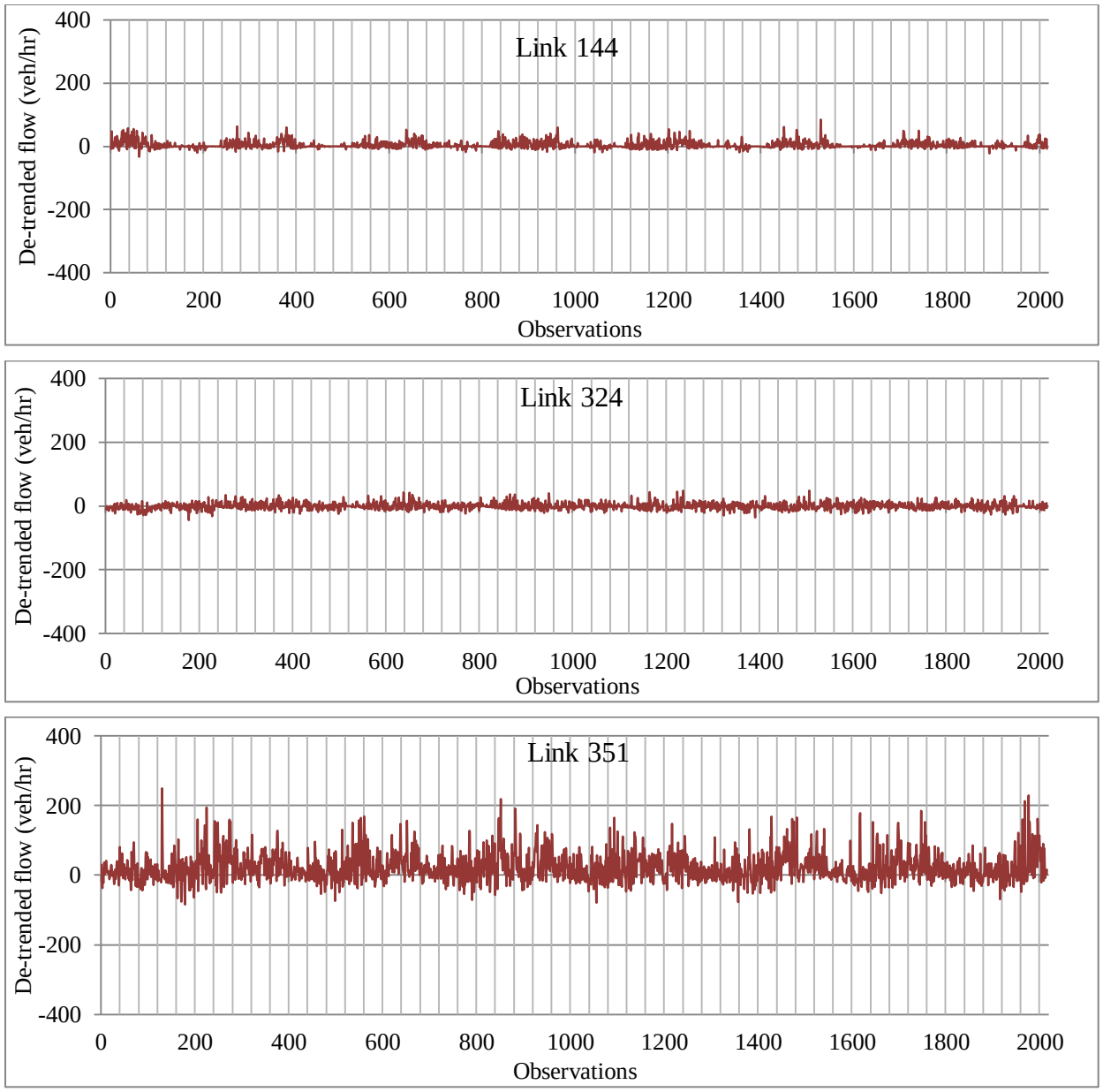
We illustrate our findings by showing the de-trended series patterns for the same set of links presented in Figure 2.11 (i.e., target link, three GC links and three non-GC links). Figure 2.13 presents the de-trended time series patterns as average over one week (2nd May to 8th May, 2016) for each 5-minute interval of the day. Looking at the pattern for the target link in Figure 2.13(a), it is evident that some observations have significant fluctuations, even after de-trending the flow data, possibly because of incidents or non-recurring traffic congestion. Looking at the patterns for the GC links in Figure 2.13(b), fluctuations of flows are also observed and, most notably, they occurred in GC links before their occurrence in the target link. This finding appears consistent with the assumption of Granger-causal relation that 'the cause happens before the effect'. Moreover, some of these fluctuations of flow occurred in the GC link after their manifestation in the target link, given that also the target link is a causal link to the GC links. This finding suggests that the Granger-causal relation between these two links is bidirectional. Looking at the patterns for non-GC links in Figure 2.13(c), none of the sharp increases or decreases of traffic flow values are observed in the de-trended time series, consistently with the non-causal relation between the links.



(a) Target link



(b) GC links of the target link



(c) Non-GC links of the target link

Figure 2.13: De-trended traffic flow time series of a target link, GC links and non-GC links.

2.4.5 Spatial characteristics of GC links

The proposed spatial variable selection method not only identifies the set of the most relevant GC links according to their GC-strength, but also finds the optimal time lag for each predictor by comparing the BIC presented in Eq. (2.5). Looking at the location of the GC links in relation to their GC-strength and optimal lag allows to understand the spatial characteristics of the GC links.

Specifically, it is interesting to observe whether distance has any relation with either the optimal time lag or the GC-strength. The analysis of the optimal lag allows to understand by looking at the relation between distance and lag.

We illustrate our findings by presenting the optimal time lag between the target link and each of the GC links for that target link. It should be noted that the time lag for which the BIC is the lowest is considered to be the optimal time lag, and its value is computed for each predictor of the target link. Figure 2.14 shows the target links and its GC links in terms of both their location and their optimal time lag, with a simplification for illustration purposes into three categories: (i) 10-14 time lag, (ii) 15-19 time lag, and (iii) 20-24 time lag. When looking at the spatial distribution of the GC links in relation to the optimal lag, it can be observed that there is no apparent relation between the distance from the target link and the GC link and the optimal time lag. In fact, it is possible that a nearer road link has a higher optimal time lag than the one of a road link more distant from the target link. Accordingly, the optimal time lag between road links depends on their time series values rather than distance between them.

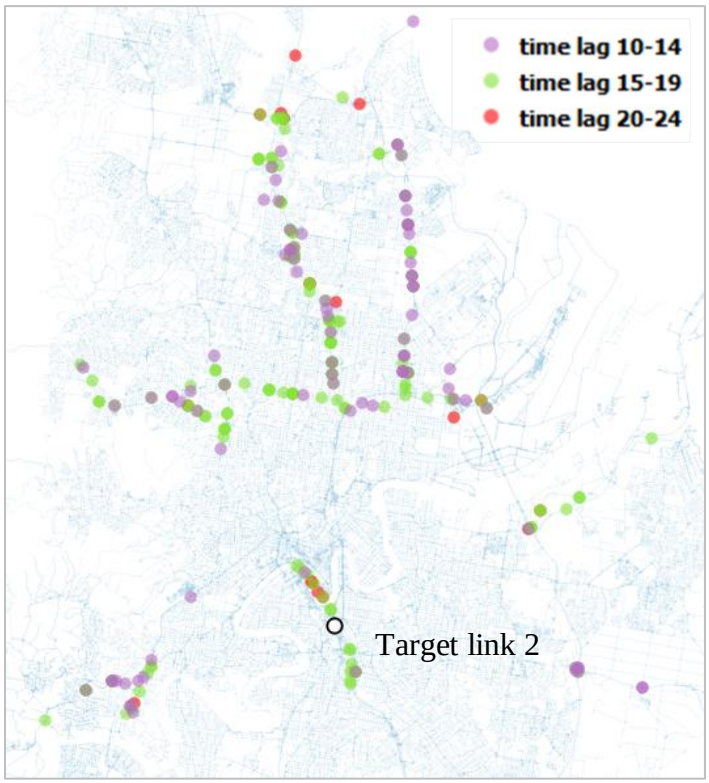


Figure 2.14: Location and optimal time lag of the target and GC links.

We also illustrate our findings by mapping the GC-strength in Figure 2.15, where we show the location and the GC-strength of the GC links for target link 2 to observe whether a relation exists between distance and GC-strength between the links. In order to simplify the illustration, three categories are presented for the GC-strength of each GC link to the target link: (i) GC-strength between $1x10^{-4}$ and $1x10^{-3}$, (ii) GC-strength between $1x10^{-3}$ and $1x10^{-2}$, and (iii) GC-strength between $1x10^{-2}$ and 1. When looking at the spatial distribution of the GC links in relation to the GC-strength, it can be noticed that there is no apparent relation between the GC-strength of the links and their distance from the target link. Also, it can be observed that the GC links with higher GC-strength are located in closer proximity to the target link 2 and, although some of the nearest links have comparatively low GC-strength, most GC links near the target link have a strong causal influence to the target link. This suggests that GC links located near the target link 2 are more useful to estimate its future traffic states of the target link 2, and, since target link 2 is a freeway link, circumstances such as congestion or incidents on a nearer link in the freeway can affect the traffic states of the target link 2 more significantly than other distant road links.

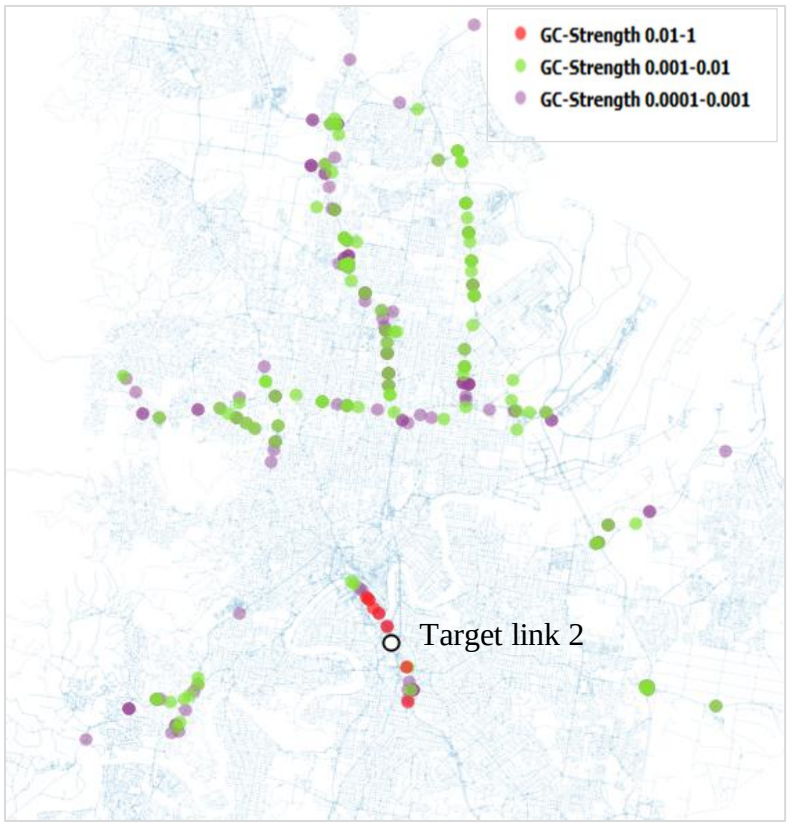


Figure 2.15: Location and GC-strength of the target and GC links.

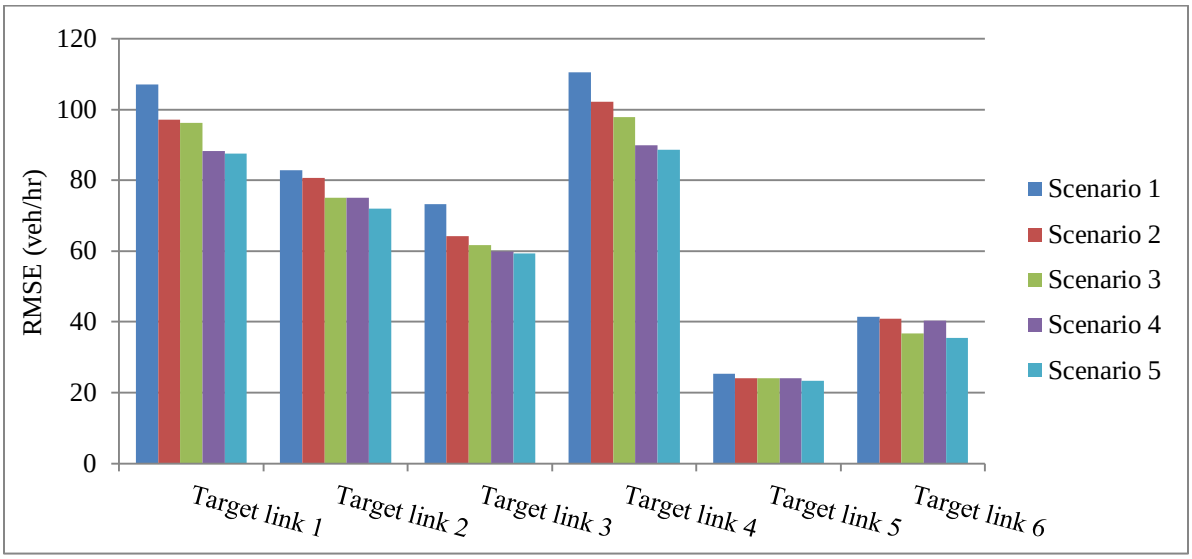
2.4.6 Prediction accuracy

In summary, we have proposed a spatial variable selection method and analysed the characteristics of the set of predictor links that guarantees the best results in terms of prediction ability and model parsimony. Given that the method aims at selecting variables, the conclusion of our analysis of the findings consists in comparing the prediction accuracy of two prediction models (i.e., time-series regression, feedforward artificial neural network) with five different scenarios (i.e., sets of predictors) as explained in section 2.3.3. Figures 2.16 and 2.17 compare the prediction errors in terms of RMSE and MAE across the five scenarios for the time-series regression model and the neural network model, respectively.

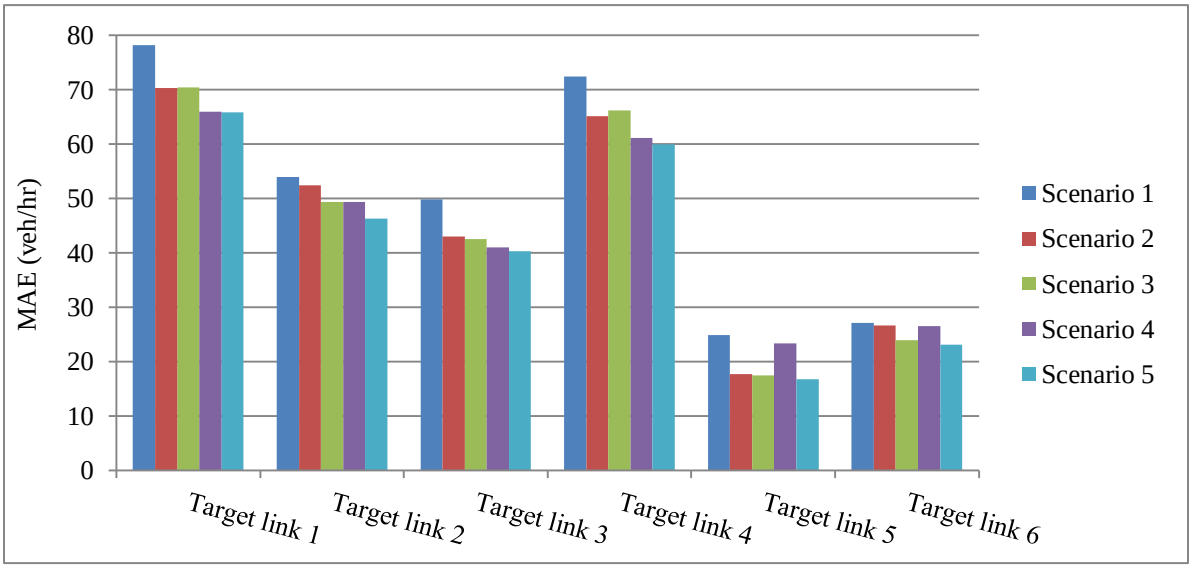
It can be observed that there is consistency in terms of performance of the different sets of predictors. Considering only the past states of the target link (scenario 1) is certainly the simplest solution, but also gives the poorest performance when it comes to prediction accuracy. Adding the moving average of past traffic states (scenario 2) contributes to the accuracy of predictions, but not sufficiently when compared to the alternative methods. Extending the set of predictor links to the nearest upstream and downstream links (scenario 3) increases the performance of both models. It should be noted that the implementation of the proposed GC-based variable selection method found for freeway links that the nearest upstream and downstream links have a higher GC-strength to the target link. On the one hand, this confirms that considering those links is beneficial to the prediction accuracy, but on the other hand this also clarifies that they should be considered perhaps only for specific road types.

Further enlarging the neighbourhood boundaries by adding a number of links equal to the one obtained with our proposed method (scenario 4) showed increased performance, with exceptions for target links 5 and 6 with the time-series regression model, and target links 3 and 4 with the neural network model. When compared in particular with the results from our proposed variable selection method (scenario 5), it appears that it is definitely not a matter of selecting a certain number of links, but rather a matter of finding a certain number of links that are significantly related to the target. Clearly, the spatial distribution of the GC links in scenario 5 shows that also links far from the target links play a significant role in predicting their traffic state, which implies

that only having larger neighbourhood boundaries as in scenario 3 does not help since several non-GC links are included.



(a) RMSE

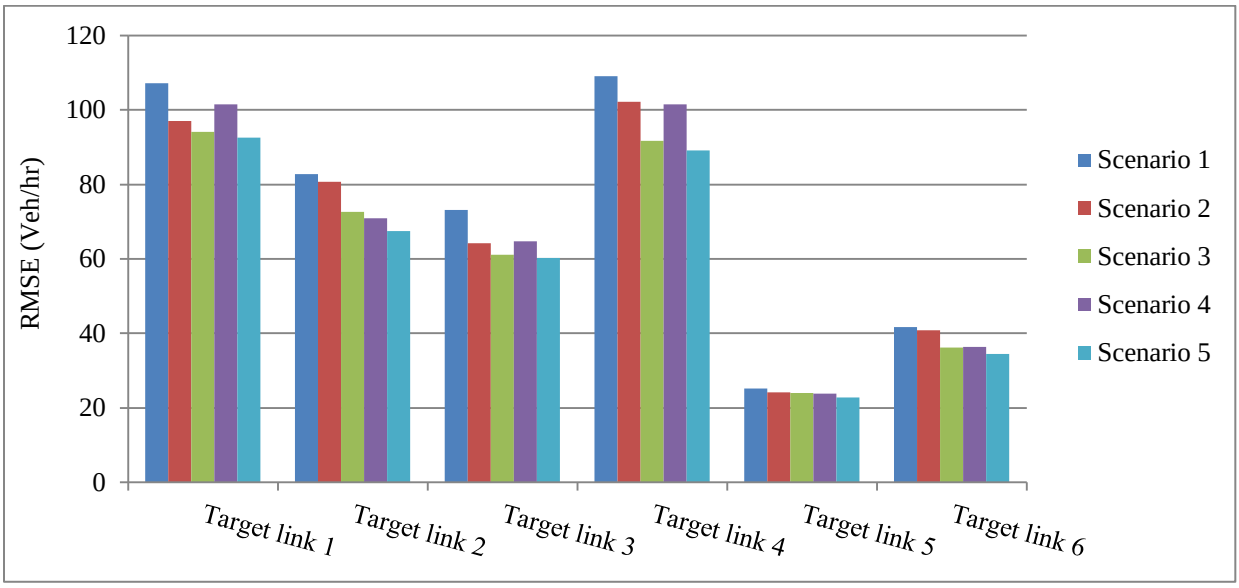


(b) MAE

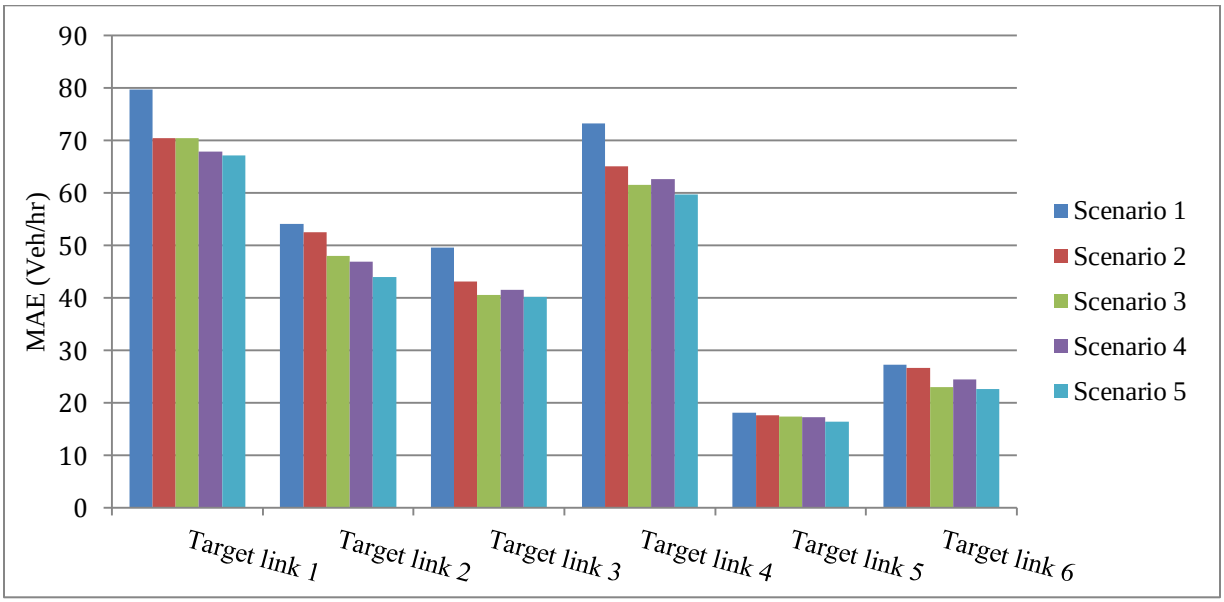
Figure 2.16: Prediction errors of the linear regression models across the five scenarios of predictor links.

Ultimately, the most important observation is that the proposed method leads to the highest prediction accuracy regardless of the target link and the prediction model. In fact, the findings

from Figure 2.17 lead to the same considerations from Figure 2.16: this is a crucial finding because it shows that the selection of predictors by pushing the spatial boundaries to the entire network gives a set of predictor links regardless of the actual model used for the prediction itself.



(a) RMSE



(b) MAE

Figure 2.17: Prediction errors of the multi-layer perceptron across the five scenarios of predictor links.

Therefore, considering the outcomes of the tested short-term traffic prediction models, GC-strength is an effective criterion to select more relevant predictor links for each target link. As mentioned, our motivation for applying bivariate Granger causality test in the spatial variable selection method is to select the predictor set which can render higher prediction accuracy by using minimum number of predictor links. The proposed method using the percentile-based GC-strength threshold allows us to achieve this goal, offering a quantitative method to select the strongest GC links.

2.5 Conclusion

This study proposes a method to select the most relevant predictors of the traffic state of a target link in a road network by applying the concept of the Granger causality. First, the method finds predictor links that have causal relation with the target link based on the Granger causality test, called GC links. Then, the method ranks these links according to the measure of the strength of the causal relation, known as GC-strength. Finally, the method selects the number of predictor links within the set of GC links according to this GC-strength by applying a 90% percentile threshold. The proposed method is therefore efficient in removing a significant number of irrelevant predictor links and at the same time providing a parsimonious set of causally relevant predictor links for each target link. The set of the most relevant GC links can then be used with any prediction model (e.g., time-series regression, artificial neural network), thus indicating the independence of the selection method from the prediction model itself. The parsimonious set of relevant predictor links can therefore be a very useful tool for traffic management and control scheme in the road network.

The implementation of the proposed method to the Brisbane network case-study revealed that the method is effective in finding a spatially diverse set of relevant predictor links. In fact, while existing methods mainly look at neighbouring links (when not considering only the past traffic states of the target link or simply the upstream and downstream links), the proposed method reveals that the relevant GC predictors are found across the entire network. From a spatial perspective, the distance from the target link does not seem to have a relation with the optimal time lag but appears related to the GC-strength as, at least for freeway links, upstream and

downstream links show higher values of causal relation strength. From a temporal perspective, the time of day and day-to-day fluctuations of GC links recall the patterns of the target link. Overall, the method proves effective in finding strong relations between links that, de facto, would not be considered for predictive models because far from the target link, even if their temporal profiles (even after de-trending) are similar.

Overall, this study shows that the bivariate Granger causal relation between road links is a powerful tool to select relevant predictors. Further research avenues could look into multivariate approaches to selecting GC links in a road network, investigate non-linear causal relations to relax the linear assumption of the GC test, and examine the implication of finding GC links common to all target links.

Chapter 3

IDENTIFICATION OF THE MOST INFLUENTIAL ROAD LINKS IN URBAN TRAFFIC NETWORKS

3.1 Introduction

A key element in Intelligent Transport Systems (ITS) for integrated and coordinated traffic management systems is the reliable and accurate measure and prediction of traffic states. Overall road network performance can be enhanced by anticipating properly future traffic states and thereby deploying proactively traffic control strategies to mitigate traffic congestion and provide accurate travel information to road users. One of the major concerns in estimating and predicting traffic states in large urban networks is the selection of good features that effectively capture the network-wide traffic states. Consider a network with N road links, where we wish to predict the future traffic states of these N links simultaneously. While it is possible to use all N links as predictors (i.e., predict the future states of the N links based on the past states of all the N links), it is certainly more desirable to use a smaller number of predictors if the latter contain necessary information to predict the whole system state. This is especially true when N is large because building and calibrating models with a large number of predictors can pose significant computational challenges.

In this study, we consider the problem of identifying a reduced set of predictor links that can satisfactorily support the prediction of the future traffic states on all the links in the network. One approach to this problem is to measure the *importance* of an individual link as a network-wide predictor and select the most influential predictor links. This study proposes statistical methods that measure and rank link importance in terms of how useful a given link is in predicting the traffic states of other target links. The problem is solved via a two-step procedure, where we first identify a set of relevant predictor links for each target link and then we rank all predictor links according to how often a given link is selected as relevant predictor for other target links, or equivalently how many target links are dependent on the given predictor link. The goal of the proposed methods is to enable the development of network-wide traffic models for parsimonious

and efficient prediction, which requires fewer parameters and thus significantly reduces computational costs while achieving and maintaining the desired level of prediction accuracy.

There exist several studies on variable selection methods for traffic prediction models. In regard to detecting relevant predictors for a given target link, however, most of the existing studies have employed manual approaches to specifying the predictor set. For instance, studies considered a fixed spatial boundary limited to the target link itself (Ahmed and Cook, 1979; Clark, 2003; Smith et al., 2002; Okutani and Stephanedes, 1984), the nearest upstream and downstream of the target link (Hobeika and Kim, 1994; Sun et al., 2006; Stathopoulos and Karlaftis, 2003; Chandra and Al-Deek, 2009) or nearer neighbour links (Kamarianakis and Prastacos, 2003; 2004) as the relevant predictor locations for estimating the target link's future traffic states. Overall, the traffic states of the road links that are not in close proximity to the target link were largely neglected in existing studies. In real networks, however, the traffic states of a target link can be influenced by not only the traffic states of its adjacent road links but also that of distant road links (Xu et al., 2016; Sun and Zhang, 2007). So it is imperative to use the influence of distant links on the target link in traffic forecast.

Since the effects of predictor links on the traffic states of the target link varies, there is a need to measure the importance of predictor links, namely the degree to which predictor links contribute to the prediction of the target link's future conditions. Recognising this need, some recent studies utilised the importance of predictor links to a given target link as the basis of selecting relevant predictor set. In developing short-term traffic prediction model using Gradient Boosting Regression Trees, Yang et al. (2017) employed the number of times a variable was selected for splitting the tree as the criterion for measuring importance of the predictors. Similarly, Hou et al. (2015) developed a Random Forest-based short-term prediction model and proposed improvement of node purity (or splitting of trees) as the basis of predictor link ranking. Sun and Zhang (2007) estimated Pearson correlation coefficients among the target link and each of the predictor links in the network to evaluate importance. Xu et al. (2015) applied Bayesian multivariate adaptive-regression splines to predict short-term traffic and used the frequency of selection of each variable on different samples as the criterion for importance estimation.

The previous studies generally consider a few target links and predictor links in a small road network and there are still significant research gaps in developing variable selection techniques for large-scale and complex urban road networks. Furthermore, the existing variable selection techniques focus on selecting a separate set of predictor links for each of the target links, without offering the solution on how to identify a set of predictor links for the whole network system consisting hundreds or even thousands of target links.

To address this gap, this study proposes a systematic method of identifying important predictor road links for the network-wide traffic prediction by using spatial relations among road links in the network and ranking them by their importance as predictors. Unlike existing studies, this study does not consider any fixed spatial boundary to select the predictor locations for a given target link and does not propose separate predictor sets for each of the target links. The proposed method initially considers all road links in a network as possible predictor locations and selects the relevant ones for a given target link by means of rigorous statistical methods such as *Granger causality* analysis (Granger, 1969; 1980) or *elastic net* regularisation (Zou and Hastie, 2005). These relevant predictors are then ranked according to their influence in the network and, finally, a common set of important predictor links for all the target links in the network is determined by constructing a hierarchy of importance. To the best of our knowledge, the proposed method of selecting a common set of important predictors in a large-scale urban road network is a novel approach in the traffic modelling context.

The remainder of the chapter is organised as follows. Section 3.2 describes the method of selecting a set of the most important predictor links in a road network. Section 3.3 illustrates the case-study and details the application of the methods to the large-scale network of Brisbane. Last, Section 3.4 illustrates and discusses the results of the case-study followed by Section 3.5 that summarises the conclusions of this study.

3.2 Detecting important predictor links using spatial variable selection techniques

This section presents the methods for selecting a common set of important links in the road network by using a two-step procedure: (i) the first step uses a spatial variable selection technique to identify the spatially related predictor road links for each target road link; (ii) the second step identifies a common set of best predictor links for all the target links. In the first step, two spatial variable selection techniques are considered: *Granger causality* (GC) analysis and *elastic net* (EN) regularisation. The GC analysis employs *F*-statistics in Vector Auto Regression (VAR) to select predictors by using ordinary least square method for residual estimation. The EN uses a regularisation technique with optimisation-based solution to select predictors by using penalised least square methods for residual estimation. In the second step, the *k* most influential road links are selected as a common predictor set for the whole network, where the importance of a link is evaluated by the percentage of times that it was selected as a relevant predictor link by a spatial variable selection technique (i.e. GC or EN) in the first step.

3.2.1 Variable selection using Granger causality analysis

The GC analysis recognises directed functional or causal interactions of different variables in time series data and excludes variables without an interaction (Seth et al., 2015; Li et al., 2015). The GC analysis measures the effect of the historic values of a variable (e.g., x_{t-1}) in predicting another variable (e.g., y_t) by comparing the prediction errors before and after the inclusion of x_{t-1}. If the prediction error is reduced significantly, then it is considered to be an improvement of the prediction of y_t due to the inclusion of x_{t-1} and it is said that time series x "Granger-causes" time series y (Granger, 1969). In the traffic analysis context, the GC analysis allows to find out the spatial relations between road links in a road network. Unravelling these relations is then used to identify relevant road links that can be used to predict the traffic parameters of a given target road link.

The GC analysis can have two forms: pairwise GC and conditional GC. Pairwise GC is determined by bivariate time series analysis and conditional GC can be determined by VAR-based time series analysis. Conditional GC can identify whether a Granger-causal relation

between two time series is direct or mediated by another time series. However, this type of causal relation cannot be detected by pairwise GC. For instance, if the Granger-causal influence of x_{t-1} on y_t is entirely mediated by another variable z_{t-1}, then no significant improvement in prediction of y_t can be observed by using x_{t-1} as a predictor variable. In this case, conditional GC suggests that *x* does not Granger-cause *y*, whereas pairwise GC may detect causality of *x* on *y* (Ding et al., 2006; Dhamala et al., 2008). Due to the superiority of conditional GC over pairwise GC in detecting potentially intertwined causal influence among multiple predictors, this study adopts VAR-based GC to identify relevant predictor links for a given target link in the road network.

In a VAR model, each variable is taken as response variable once and the other variables are considered as predictors. Multivariate time series analysis is used in a VAR model where the value of each variable at time *t* is computed by its own lagged values and the lagged values of other predictors (Zivot and Wang, 2006). Let $\boldsymbol{\theta}_t = (\theta_{1,t}, \theta_{2,t}, \cdots, \theta_{N,t})$ be an $(N \times 1)$ vector representing a traffic flow measure for *N* road links at time *t*, where $t = 1, 2, \ldots T$. The basic *P*-lag VAR model for predicting $\boldsymbol{\theta}_t$ can be written as follows:

$$\boldsymbol{\theta}_t = \mathbf{c} + \boldsymbol{\Pi}^{(1)}\boldsymbol{\theta}_{t-1} + \boldsymbol{\Pi}^{(2)}\boldsymbol{\theta}_{t-2} + \cdots + \boldsymbol{\Pi}^{(p)}\boldsymbol{\theta}_{t-p} + \cdots + \boldsymbol{\Pi}^{(P)}\boldsymbol{\theta}_{t-P} + \boldsymbol{\varepsilon}_t \quad (3.1)$$

where $\mathbf{c}$ is an $(N \times 1)$ vector of intercepts and $\boldsymbol{\Pi}^{(p)}$ is an $(N \times N)$ coefficient matrix reflecting the relationships between the response variable ($\boldsymbol{\theta}_t$) and its *p*-lagged variable ($\boldsymbol{\theta}_{t-p}$) with time lag $p = 1, 2, \cdots, P$. Eq. (3.1) can be decomposed to show the components related to a single target link n ($n = 1, 2, \cdots, N$) as follows:

$$\theta_{n,t} = c_n + \left(\pi_{n,1}^{(1)}\,\theta_{1,t-1} + \pi_{n,2}^{(1)}\,\theta_{2,t-1} + \cdots + \pi_{n,N}^{(1)}\,\theta_{N,t-1}\right) + \left(\pi_{n,1}^{(2)}\,\theta_{1,t-2} + \pi_{n,2}^{(2)}\,\theta_{2,t-2} + \cdots + \pi_{n,N}^{(2)}\,\theta_{N,t-2}\right) + \cdots + \left(\pi_{n,1}^{(P)}\,\theta_{1,t-P} + \pi_{n,2}^{(P)}\,\theta_{2,t-P} + \cdots + \pi_{n,N}^{(P)}\,\theta_{N,t-P}\right) + \varepsilon_{n,t} \quad (3.2)$$

where $\theta_{n,t}$ denotes a scalar at the n^{th} element of $\boldsymbol{\theta}_t$ representing the traffic flow measure for link n at time t and $\pi_{n,m}^{(p)}$ is a scalar at the n^{th} row and the m^{th} column of $\boldsymbol{\Pi}^{(p)}$ representing the relation between the traffic condition on link n at time t (target link or response variable) and the traffic condition on predictor link m at time $t - p$, where $n, m = 1, 2, \cdots, N$.

To determine the time lag order P, model selection criteria such as the Akaike Information Criterion (AIC) and the Schwarz-Bayesian Information Criterion (BIC) can be used. At first, the VAR model is fitted with each of the lag orders $P = 0, 1, \cdots, P_{max}$ and the corresponding value of the model selection criterion is calculated. Then, the actual time lag order can be identified by comparing the scores of the AIC and BIC:

$$\text{AIC} = -2\ln(L) + 2q \tag{3.3}$$

$$\text{BIC} = -2\ln(L) + q\ln(T) \tag{3.4}$$

where L is the maximised value of the likelihood function of the model at the value of the parameter estimates, q is the number of parameters in the model, and T is the number of observations. It should be noted that the AIC or BIC scores decrease with an improvement in the log-likelihood and a decrease in the number of parameters. The lag order with the lowest AIC or BIC score is considered as the best lag order (Cottrell and Lucchetti, 2016). Since the main objective of this study is to obtain the most parsimonious model, the lowest lag order between AIC and BIC is selected in this study.

After the specification of the VAR model, GC between a target link and each of its predictor links is tested by performing an F-test. Consider target link n and predictor link m: a time series of predictor link $\theta_{m,t}$ is considered to Granger-cause a time series of target link $\theta_{n,t}$ if at least one of the lagged values of $\theta_{m,t}$ provides stastically significant information about future values of $\theta_{n,t}$. This can be tested through the F-test with the null hypothesis $\text{H}_0: \pi_{n,m}^{(1)} = \pi_{n,m}^{(2)} = \cdots = \pi_{n,m}^{(P)} = 0$ and the alternative hypothesis $\text{H}_1: \left(\pi_{n,m}^{(1)} \neq 0\right) \cup \left(\pi_{n,m}^{(2)} \neq 0\right) \cup \cdots \cup \left(\pi_{n,m}^{(P)} \neq 0\right)$. The null hypothesis that $\theta_{m,t}$ does not Granger-cause $\theta_{n,t}$ is rejected if at least one of the elements $\pi_{n,m}^{(p)}$ for $p = 1, 2, \cdots, P$ is significantly larger than zero (Bahadori and Liu, 2018). The F-test statistic is computed as follows:

$$F_0 = \frac{\frac{RSS_r - RSS_{ur}}{v}}{\frac{RSS_{ur}}{T-(s+1)}} \tag{3.5}$$

where RSS_r is the sum of the squared residuals of a restricted model (e.g., the model with $\pi_{n,m}^{(1)} = \pi_{n,m}^{(2)} = \cdots = \pi_{n,m}^{(P)} = 0$), RSS_{ur} is the sum of the squared residuals of an unrestricted model (e.g., the full model in Eq. (3.1)), *v* is the number of restrictions or the number of coefficients being jointly tested, T is the number of observations, and *s* is the number of explanatory variables in the unrestricted model. The F_0 value is then compared with the critical value of F at the 0.01 significance level. If the F_0 value is higher than the critical value, it rejects the null hypothesis and hence the statement that the time series of the traffic state on the tested predictor link does not Granger-causes the time series of the traffic state on the target link. Otherwise, the null hypothesis cannot be rejected, indicating that the tested predictor link does not provide statistically significant information about future states on the target link.

3.2.2 Variable selection using elastic net regularisation

In the areas of statistics and machine learning, regularisation is considered as a procedure to introduce additional information into a predictive model to prevent statistical over-fitting. It works by adding a penalty term associated with model complexity to the objective function of the optimization procedure for parameter estimation. In regression analyses, regularisation techniques are often used to remove irrelevant or redundant predictors to make the model more parsimonious and improve the prediction accuracy, especially when the model involves high dimensional data (Zou and Zhang, 2009). Typical regularisation techniques are *lasso* (least absolute shrinkage and selection operator), *ridge*, and *EN* methods.

The *lasso* method uses the L_1–norm of the regression coefficients as the penalty term (Tibshirani, 1996). This L_1–norm penalty drives some of the coefficients to be zero and thus can be used for the selection of the best predictor subset. However, the lasso method is known to have stability problems, namely the selected predictor subset may vary a lot when high correlation between predictors exists, because the lasso will only select one of the correlated predictors at random (Zou and Hastie, 2005; Zou and Zhang, 2009; Tibshirani, 1996). The *EN* method overcomes this instability issue by adding another penalty term based on the L_2–norm (Zou and Hastie, 2005; Ogutu et al., 2012). The L_2-norm penalty, which when used alone is referred to as the *ridge* method (Hoerl and Kennard, 1988), tends to shrink the coefficients equally towards zero,

encouraging the model to take into account all input variables thereby preventing over-fitting and improving prediction accuracy. By having a mixture of the L_1 (lasso) and L_2 (ridge) penalty terms, the EN method effectively performs variable selection (via lasso) without compromising the model stability and prediction accuracy (via ridge). The model prediction performance of the EN is generally better than that of the other two regularisation techniques in the context of the analysis of high dimensional data (Zou and Hastie, 2005; Zou and Zhang, 2009; Li and Lin, 2010). Therefore, EN is selected for this study and performed according to the following process.

Consider the linear regression model $\theta_{n,t} = c_n + \boldsymbol{\theta}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_n + \varepsilon_{n,t}$, where the response variable $\theta_{n,t}$ is the traffic flow measure on target link n ($n = 1, 2, \dots, N$) at time t ($t = 1, 2, \dots, T$), $\boldsymbol{\theta}_{t-1} = \left(\theta_{1,t-1}, \theta_{2,t-1}, \cdots, \theta_{N,t-1}\right)^{\mathrm{T}}$ is an $(N \times 1)$ vector of explanatory variables representing the lagged values of the traffic flow measure for all N road links including the target link (the superscript T represents the transpose operator), $\boldsymbol{\beta}_n = \left(\beta_{n,1}, \beta_{n,2}, \cdots, \beta_{n,N}\right)^{\mathrm{T}}$ is an $(N \times 1)$ coefficient vector, and c_n and $\varepsilon_{n,t}$ are respectively the constant and error terms. For any fixed non-negative penalty parameters λ_1 and λ_2, the EN loss function is defined as follows (Zou and Hastie, 2005):

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}_n) = \left\{\frac{1}{T}\sum_{t=1}^{T}\left(\theta_{n,t} - c_n - \boldsymbol{\theta}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_n\right)^2\right\} + \lambda_1\|\boldsymbol{\beta}_n\|_1 + \lambda_2\|\boldsymbol{\beta}_n\|^2 \qquad (3.6)$$

where $\|\boldsymbol{\beta}_n\|_1 = \sum_{i=1}^{N}\left|\beta_{n,i}\right|$ is the L_1-norm of the coefficients and $\|\boldsymbol{\beta}_n\|^2 = \sum_{i=1}^{N}\left|\beta_{n,i}\right|^2$ is the L_2-norm of the coefficients, representing respectively the L_1–penalty (lasso) and L_2–penalty (ridge) terms. The EN estimator $\widehat{\boldsymbol{\beta}}_n$ minimises the following equation:

$$\widehat{\boldsymbol{\beta}}_n = argmin_{\boldsymbol{\beta}_n}\{L(\lambda_1, \lambda_2, \boldsymbol{\beta}_n)\} \qquad (3.7)$$

The minimisation procedure can be written as a penalised least squares method. Let $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$ and $1 - \alpha = \frac{\lambda_1}{\lambda_1+\lambda_2}$, then Eq. (3.6) can be rewritten as:

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}_n) = \left\{\frac{1}{T}\sum_{t=1}^{T}\left(\theta_{n,t} - c_n - \boldsymbol{\theta}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_n\right)^2\right\} + (\lambda_1 + \lambda_2)[(1-\alpha)\|\boldsymbol{\beta}_n\|_1 + \alpha\|\boldsymbol{\beta}_n\|^2] \qquad (3.8)$$

Defining $\lambda = (\lambda_1 + \lambda_2)$, Eq. (3.8) becomes:

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}_n) = \left\{\frac{1}{T}\sum_{t=1}^{T}\left(\theta_{n,t} - c_n - \boldsymbol{\theta}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_n\right)^2\right\} + \lambda[(1-\alpha)\|\boldsymbol{\beta}_n\|_1 + \alpha\|\boldsymbol{\beta}_n\|^2] \quad (3.9)$$

Solving for $\widehat{\boldsymbol{\beta}}_n$ in Eq. (3.9) is equivalent to solving the following optimization problem:

$$\widehat{\boldsymbol{\beta}}_n = argmin_{\boldsymbol{\beta}_n}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\theta_{n,t} - c_n - \boldsymbol{\theta}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_n\right)^2\right\},$$

$$\text{subject to } \lambda[(1-\alpha)\|\boldsymbol{\beta}_n\|_1 + \alpha\|\boldsymbol{\beta}_n\|^2] \leq u, \text{where } u > 0 \quad (3.10)$$

The term $\lambda[(1-\alpha)\|\boldsymbol{\beta}_n\|_1 + \alpha\|\boldsymbol{\beta}_n\|^2]$ is the EN penalty, which is a convex combination of the lasso and ridge penalties (Zou and Hastie, 2005). The EN penalty is controlled by two tuning parameters, α and λ. The value of α can vary between 0 and 1, where 0 corresponds to the lasso and 1 corresponds to the ridge regression, and can be set prior to the calculation. The value of λ is selected via cross validation by initially specifying a set of values in order to avoid intensive computation (Li and Lin, 2010). Once the values of α and λ are defined, Eq. (3.10) can be solved by using a cyclical coordinate descent algorithm (Friedman et al., 2010). The cyclical coordinate descent algorithm successively optimizes the objective function over each parameter by keeping other parameters fixed, and cycles repeatedly until convergence (Hastie and Qian, 2014). The result of the optimisation process is a vector of parameter estimates $\widehat{\boldsymbol{\beta}}_n = \left(\hat{\beta}_{n,1}, \hat{\beta}_{n,2}, \cdots, \hat{\beta}_{n,N}\right)^{\mathrm{T}}$, where some estimates are reduced to zero due to the EN penalty. Predictors whose coefficients are reduced to zero can be considered as irrelevant to the response variable and therefore removed from the model, whereas relevant predictors with a reduced value of coefficients are kept in the model. With this process, the EN can provide a parsimonious model where the selected predictors are the links that have high spatial dependence with the target link and hence serve as an effective spatial variable selection technique.

3.2.3 Selection of the common set of important predictors

The spatial variable selection technique using VAR-based GC analysis or EN regularisation selects the distinct relevant links for each target link in the road network. The next step is the selection of a common set of important predictor links for all target links. Having the same predictor set for all the target links can greatly simplify the model building procedure because the same model structure can be used for all of the target links in the road network. Also,

identifying the same set of relevant predictors can be considered as a method of finding the optimal sensor locations for network-wide traffic estimation and prediction, which guides where to obtain traffic observations to best serve the prediction of the various target links in the network as a whole.

The proposed method of selecting a common set of important predictor links for all target links is based on the importance of the predictor links in the road network. In this study, the importance of a predictor links is evaluated by the percentage of times that it was selected as the relevant predictor link in a spatial variable selection technique (i.e., GC or EN). The more times a link is selected as the relevant predictor for a target link, the more important this link is because this link serves as a predictor for more target links. It is also possible that a link is not selected by any of the target link, meaning that no target link finds the link to be relevant and it is not considered as an important predictor.

Once we know the importance of each link in terms of how many target links are dependent on this link, we can construct a hierarchy according to their importance. Therefore, the common set of important predictors for a network can be identified by selecting the k important predictors. The choice of k depends on the desired level of prediction accuracy, computational cost, and budget (e.g, purchase and maintenance cost of loop detectors). In the case study, we compare different k values to investigate their effect k on network-wide short-term traffic prediction applications.

3.2.4 Evaluation

In order to assess the efficacy of the proposed network-wide variable selection method, a simple short-term traffic prediction experiment is designed. Given N road links, we first identify k common predictor links for the entire network using the proposed method and then predict the traffic state of each of the N links using the same k predictors to evaluate the prediction accuracy.

For the short-term traffic prediction model, we consider a simple time series regression that has the traffic flow on the target link as the response variable and the traffic flow on the common set of important predictor links at time lag 1 as explanatory variables. The prediction accuracy of

each prediction model for a particular target link n is measured using the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE) as follows:

$$RMSE_n = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\theta_{n,t} - \hat{\theta}_{n,t}\right)^2} \quad (3.11)$$

$$MAE_n = \frac{1}{T}\sum_{t=1}^{T}\left|\left(\theta_{n,t} - \hat{\theta}_{n,t}\right)\right| \quad (3.12)$$

where T is the number of observations, $\theta_{n,t}$ is the observed value of a traffic state of target link n at time t, and $\hat{\theta}_{n,t}$ is the associated model predicted value. To evaluate the overall performance of the network-wide traffic prediction, we also compute the average of $RMSE_n$ and MAE_n over all target links $n = 1, 2, \ldots, N$ as a summary measure as follows:

$$Mean\ RMSE = \frac{1}{N}\sum_{n=1}^{N} RMSE_n \quad (3.13)$$

$$Mean\ MAE = \frac{1}{N}\sum_{n=1}^{N} MAE_n \quad (3.14)$$

3.3 Case study

For this case study, the road network and the traffic state data as described in section 2.3.1 and section 2.3.2 are nominated. The traffic state data of each of the 479 road links in the road network are considered as a target link once and the traffic states of the remaining road links are selected as predictor links.

3.3.1 Spatial variable selection

Considering each of the 479 links as a target link, we applied the two variable selection techniques described in Section 3.2 to identify reduced predictor links for the given target link. For the variable selection techniques (i.e., for Eq. (3.1) and Eq. (3.10)), lag order 1 was selected on the basis of the BIC values. The GC analysis was implemented by using the 'Gretl' software package (Cottrell and Lucchetti, 2016). The EN regularisation was applied by using, the 'glmnet' software package in R (Hastie and Qian, 2014). The EN required the two tuning parameters, α and λ, to be predefined as shown in Eq. (3.10). For α, the value of 0.5 was used to weigh equally

the lasso and ridge penalties. For λ, the lowest and highest values in its range were considered, where λ_{min} represents the lowest λ at which the minimum MSE is obtained in Eq. (3.10), and λ_{1se} represents the highest λ at which the MSE attained in Eq. (3.10) is within one standard error (*1se*) of the minimum MSE. Henceforth, these two approaches of EN method are named as EN(λ_{min}) and EN(λ_{1se}). In regards to selecting a common set of predictor links for the whole network, namely the *k* most important predictors in the network, we considered five values of *k*, namely 10, 20, 30, 40, and 50.

3.3.2 Performance analysis

To evaluate the effectiveness of the proposed network-wide predictor selection methods, we considered four different scenarios of specifying input variables for short-term prediction models as follows:

1) *Scenario 1* (GC): a prediction model includes the *k* most important predictor links identified by the GC method.
2) *Scenario 2* (EN($\lambda_{\min}$)): a prediction model includes the *k* most important predictor links identified by the EN($\lambda_{\min}$) approach.
3) *Scenario 3* (EN(λ_{1se})): a prediction model includes the *k* most important predictor links identified by the EN(λ_{1se}) approach.
4) *Scenario 4* (Random): a prediction model includes *k* randomly selected links in the network.

Given each value of *k*, four different prediction models were built for each of the 479 links based on these four scenarios. The model specifications differ only in the predictor set and all the models used the lag order of 1.

Among the aforementioned four scenarios, Scenarios 1-3 represent the methods of choosing a common predictor set based on the statistical approaches proposed in this study (i.e., GC and EN), whereas Scenario 4 was introduced to compare the performance of these three methods to the case where no statistical dependence is considered. To reduce potential variations in the comparison results due to the randomness in Scenario 4, we prepared 10 different sets of randomly selected predictor links and computed the average prediction performances from these

10 prediction results as the prediction performance of Scenario 4.

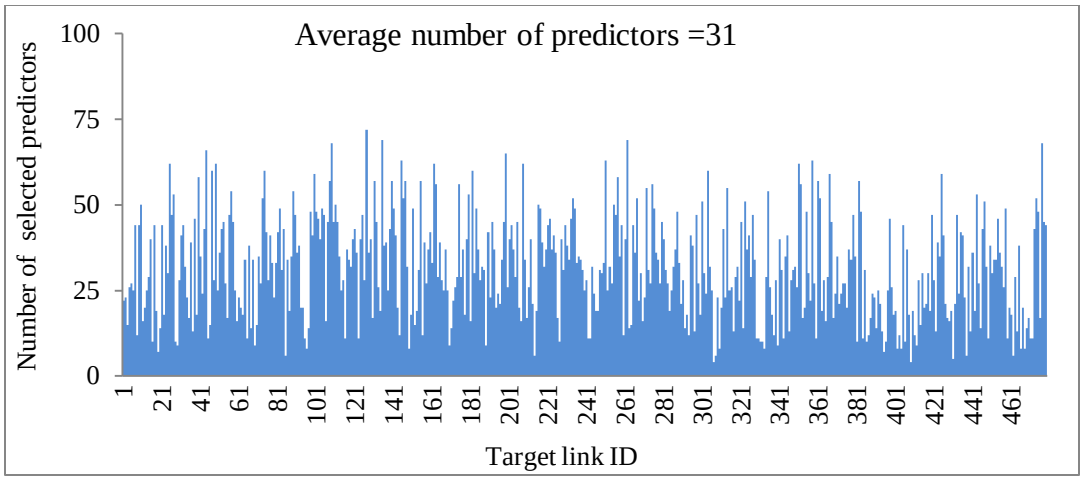
With each of the 479 links as a target link and considering the four scenarios, a total of 6,227 short-term traffic prediction models were built and tested for each *k* value (13 models for each target link). The model parameters were estimated using 80% of the data (training set) and prediction performance are evaluated using the remaining 20% of the data (testing set).

3.4 Results and discussions

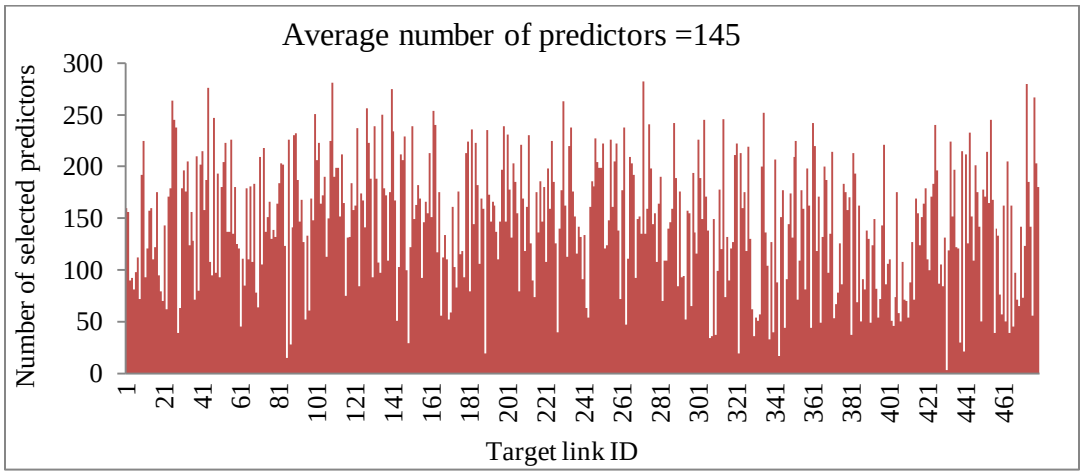
This section discusses the performance of the proposed methods of selecting the common set of *k* most important predictors in terms of their ability to reduce the model dimensionality (and hence reduce the model complexity and computational costs) and the ability to enhance the prediction accuracy through the removal of irrelevant predictors.

3.4.1 Dimensionality reduction

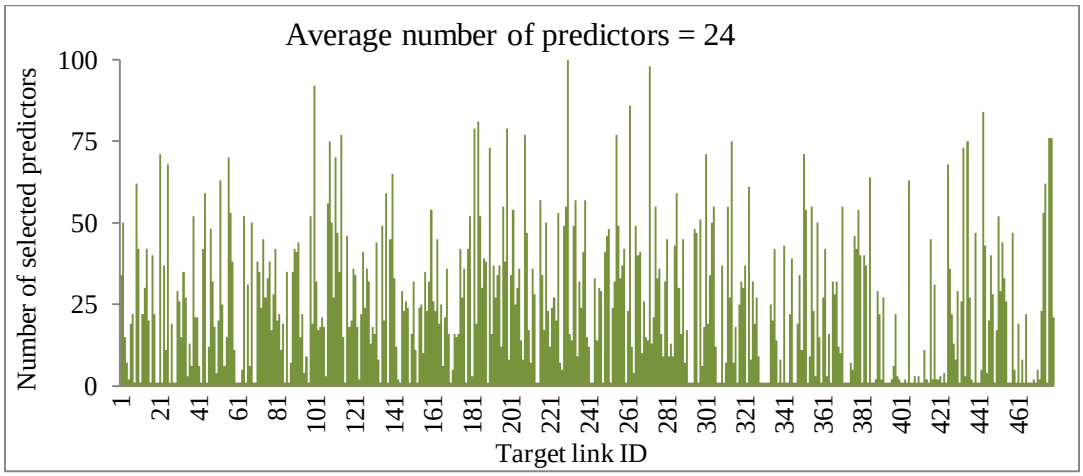
The variable selection techniques GC, EN(λ_{min}), and EN(λ_{1se}) are efficient to reduce the non-significant predictors and select only the relevant ones for a given target link. Figure 3.1 shows the total number of relevant predictor links selected by each of the three variable selection techniques with respect to all 479 target links. Overall, GC and EN(λ_{1se}) reduce the number of predictors significantly from potential 479 links to, on average, 31 and 24 predictor links, respectively. The dimensionality reduction effect is relatively lower in EN(λ_{min}), compared to GC and EN(λ_{1se}), producing 145 predictors on average. This is because EN(λ_{min}) drives more heavily towards minimizing the prediction error (the minimum MSE) and, thus, requires more predictors to improve the accuracy. EN(λ_{1se}), on the other hand, allows higher prediction error than EN(λ_{min}), thereby leading to less predictors than EN(λ_{min}).



(a) GC



(b) EN(λ_{min})



(c) EN(λ_{1se})

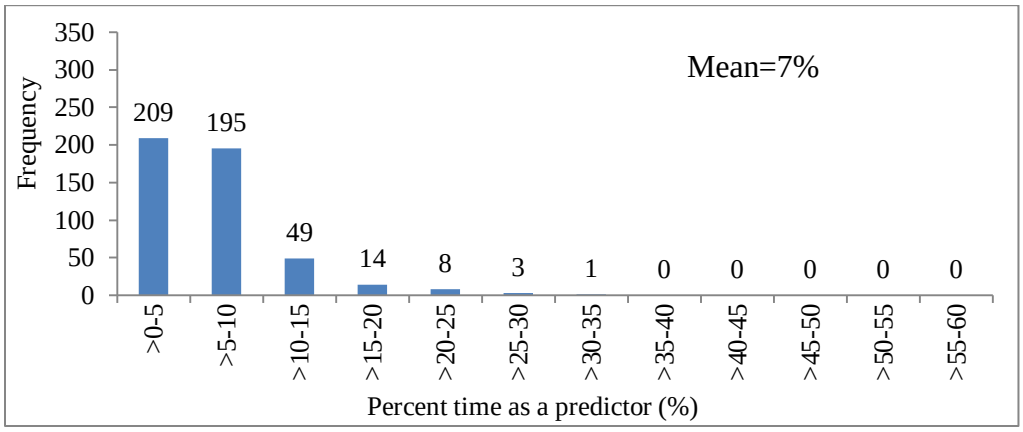
Figure 3.1: Number of predictors selected for each target link by three variable selection techniques: GC, EN(λ_{min}), and EN(λ_{1se}).

The importance of each link is assessed as the percentage of times that each link was selected as the relevant predictor. Since we have a total of 479 target links, each link can be selected as a relevant predictor up to a maximum of 479 times, which happens when the link is selected as the predictor for every target link, indicating that every link in the network is dependent on this link. It is also possible that a link is not selected by any of the target link, meaning that no target link finds the link to be relevant. The spatial variable selection techniques GC, EN($\lambda_{\min}$) and EN(λ_{1se}) identified statistically relevant predictor links to each target link, and the obtained relevant predictor links of each target links were different. It was found that a predictor link is selected as the relevant link by at least 0.84% (4) of the target links in the GC analysis, 11.90% (57) of the target links in the EN($\lambda_{\min}$) regularisation, and 0.20% (1) target links in the EN(λ_{1se}) regularisation. This indicates that every predictor has an influence on the network in terms of prediction.

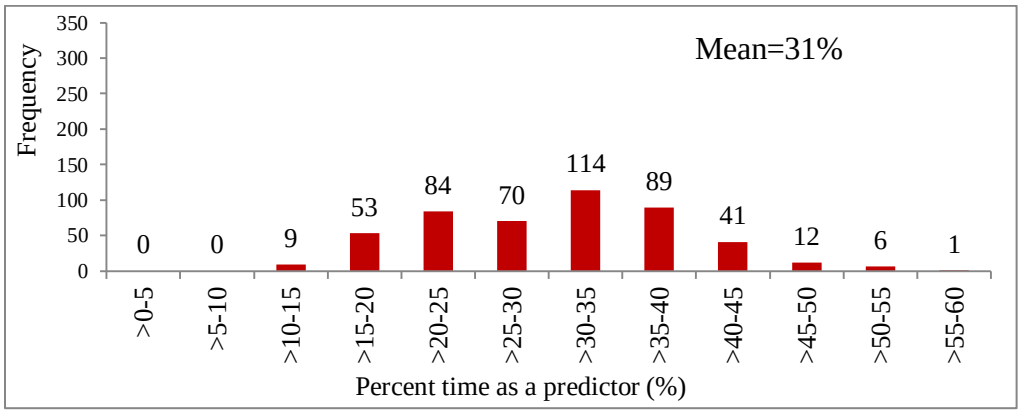
We divided the selected predictor links according to their importance into 12 categories: >0-5%, >5-10%, >10-15%, >15-20%, >20-25%, >25-30%, >30-35%, >35-40%, >40-45%, >45-50%, >50-55%, >55-60%, showing the percentage of times that a link was selected as a predictor by each spatial variable selection techniques. For instance, >0-5% indicates that a link is selected as a relevant predictor less than 24 times ($0.05 \times 479 \cong 24$), namely less than 24 target links out of the 479 target links are dependent on this link, and >55% indicates that a link was selected more than 263 times ($0.55 \times 479 \cong 263$), namely more than 263 target links are dependent on this link. The higher the percentage is, the more important a link is as a network-wide predictor.

Figure 3.2 shows the distribution of the links across these 12 categories. With the GC analysis, the links were selected as a predictor for an average of 7% of the target links, whereas the links were selected for an average of 31% and 5% of the target links with the EN($\lambda_{\min}$) and EN(λ_{1se}) regularisation, respectively. Looking at the distribution of link importance within each of the spatial variable selection techniques, it is observed that some links are significantly more important than others. For instance, in the case of GC and EN(λ_{1se}), the distribution of the importance of predictors is right-skewed and there are a small number of links that serve as the predictors for more than 20% of the target links, while most links are selected for less than 10% of the target links. On the other hand, in the case of EN($\lambda_{\min}$), the distribution of importance of

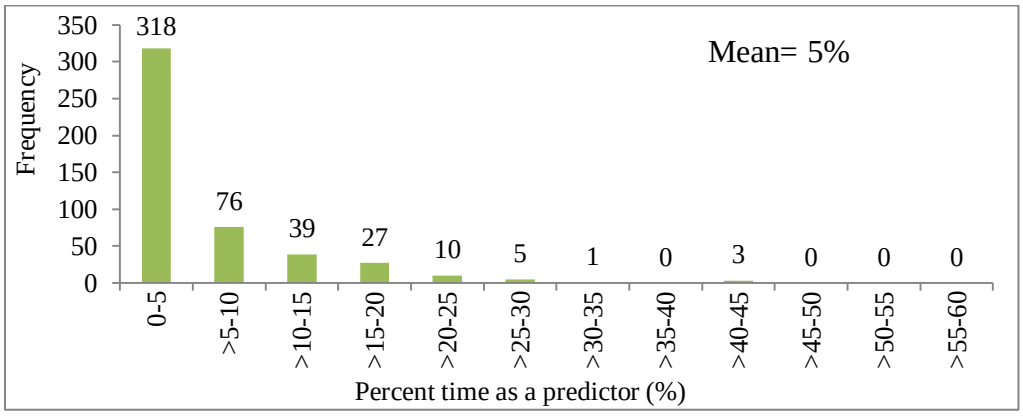
predictors is quite normally distributed where most of the links are selected for 20%-40% of the target links. Since GC and EN(λ_{1se}) select a significantly lower number of predictors than EN(λ_{min}), each link is associated with a lower number of target links when GC or EN(λ_{1se}) is used than when EN(λ_{min}) is used.



(a) GC



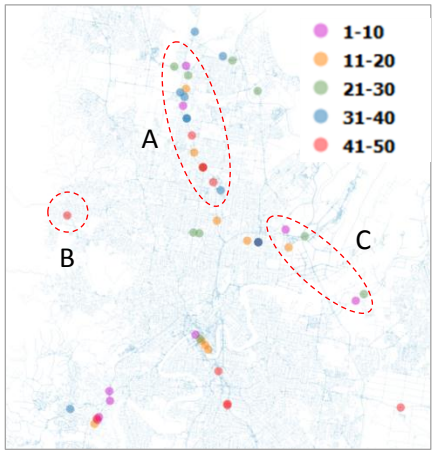
(b) EN(λ_{min})



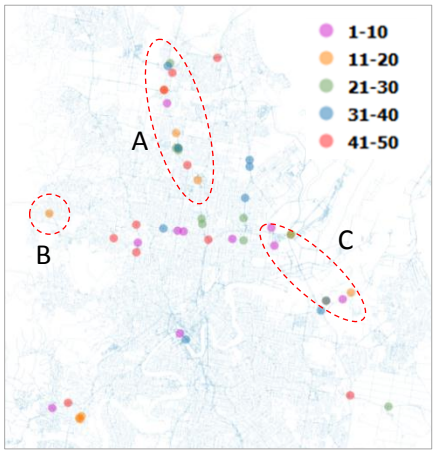
(c) EN(λ_{1se})

Figure 3.2: Distribution of number of times (in percentage) the links are selected as the relevant predictors in GC, EN(λ_{min}), and EN(λ_{1se}) methods.

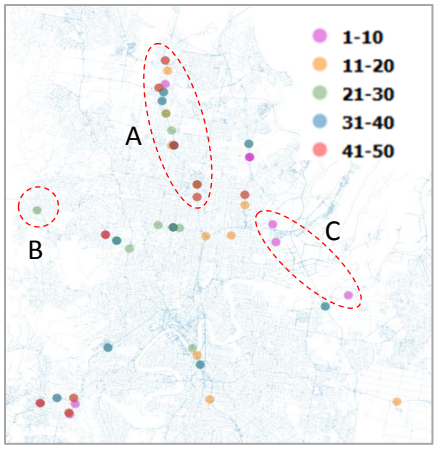
Given these measures of link importance, it is now possible to select the k most important predictors. Figure 3.3 shows the locations of the 50 most important predictors identified by GC, EN(λ_{min}), and EN(λ_{1se}). Each set of the selected 50 predictors is displayed in five categories that represent rank ranges: 1-10, 11-20, 21-30, 31-40 and 41-50. Comparing the results across GC, EN(λ_{min}), and EN(λ_{1se}), it is observed that the importance of a link as a predictor varies with the spatial variable selection techniques. There are, however, notable similarities in the three sets of selected top-50 predictors: (i) a significant number of important predictors are concentrated on the corridor in region A in Figure 3.3; (ii) the single link in region B appears in all three cases since it is noted to be an useful predictor link to predict a number of target links in the network by each method; and (iii) the similar group of 1-10 category links, which represent the top-10 important predictors, appears in region C. These observations suggest that, despite some variations, there are links that can be widely accepted as more influential than others, which can be readily captured by general time series analyses as introduced in this study.



a) GC



b) EN(λ_{min})



c) EN(λ_{1se})

Figure 3.3: Spatial distribution of the 50 most important road links selected by GC, EN(λ_{min}), and EN(λ_{1se}) methods.

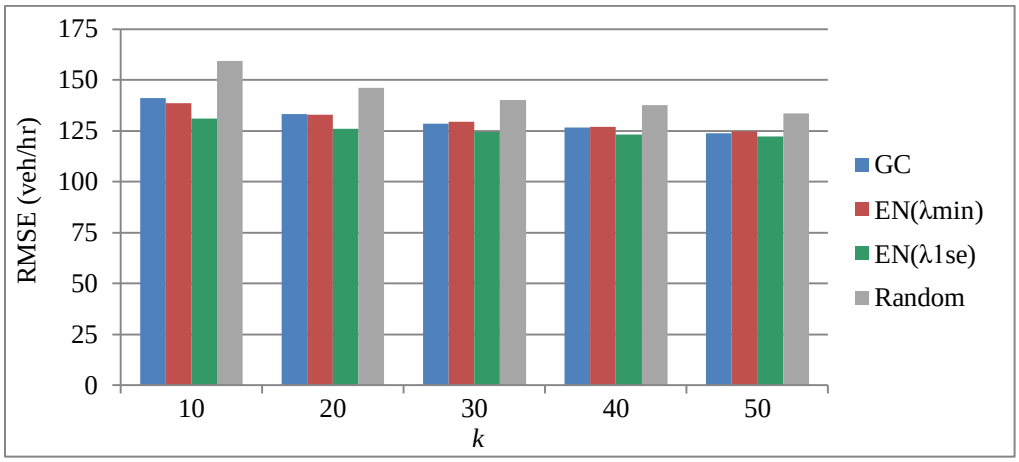
3.4.2 Prediction accuracy

Next, we evaluated the benefits of the proposed methods by analysing the ability of the selected common sets of important predictors to predict the traffic states of the target links for short-term traffic prediction. Figure 3.4 and Figure 3.5 compare the mean and the 90th percentile of RMSE and MAE of the short-term traffic prediction models with the *k* most important predictors, across different *k* values and the different predictor selection scenarios (i.e., GC, EN(λ_{min}), EN(λ_{1se}), Random) defined in Section 3.3.2. The mean and the 90th percentile are chosen as a measure of central tendency and dispersion, respectively, to analyse the performance in terms of both the mean and variation of RMSE and MAE across the target links. For all scenarios, the mean and 90th percentile RMSE and MAE decrease as *k* (the number of predictors) increases.

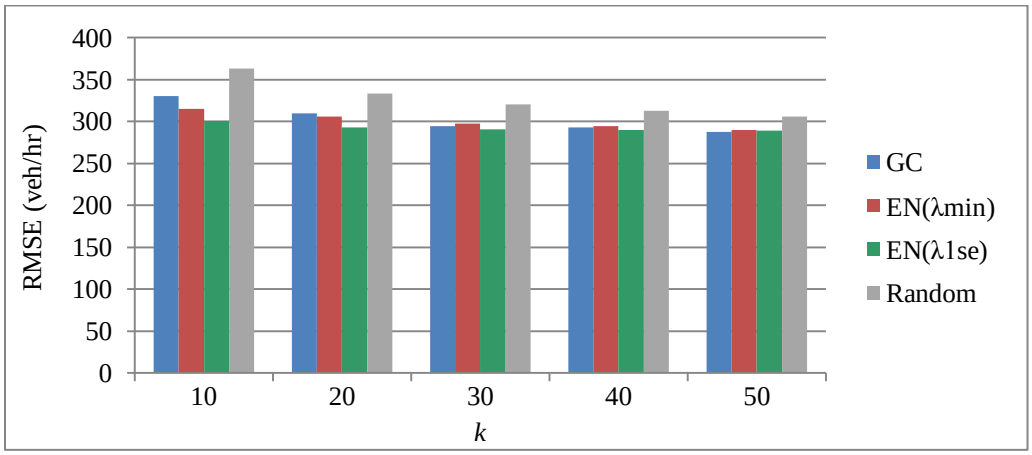
For all of the *k* values, the proposed variable selection methods (GC, EN(λ_{min}), EN(λ_{1se})) perform better than Random by yielding smaller RMSE and MAE values (mean and 90th percentile) with clear gaps between these statistical methods and the random selection. Figure 3.6 shows boxplots (mean, minimum, and maximum value) of the RMSE and MAE across the 10 sets of *k* randomly selected predictors under Scenario 4 (Random). Although there are variations in RMSE and MAE in Random, the minimum RMSE and MAE values of Random are still higher than the mean RMSE and MAE of GC, EN(λ_{min}), and EN(λ_{1se}), confirming that the

results in Figures 3.4 and 3.5 remain valid. These results imply that there exists a systematic pattern in link importance and the variable selection for traffic prediction models can be guided by this knowledge of link importance.

Among GC, EN(λ_{min}), and EN(λ_{1se}), EN(λ_{1se}) produces the best performance by showing the lowest RMSE and MAE mean and variation across nearly all of the *k* values.. The performance enhancements in the mean and the 90th percentile of RMSE and MAE are more apparent in smaller *k* values, e.g., when the prediction is done by using the top-10 and top-20 predictors.

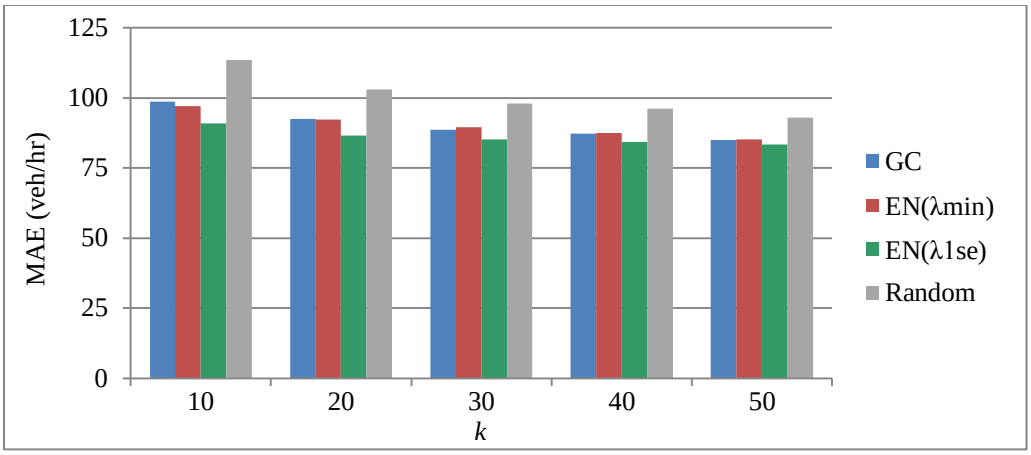


a) Mean

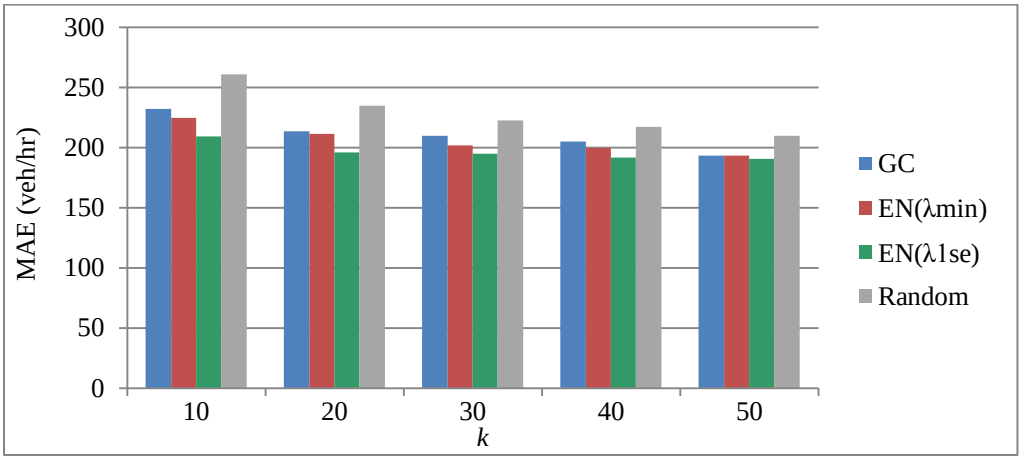


b) 90th percentile

Figure 3.4: Prediction accuracy (RMSE) of short-term prediction models with top-*k* important predictors based on the four predictor selection scenarios.

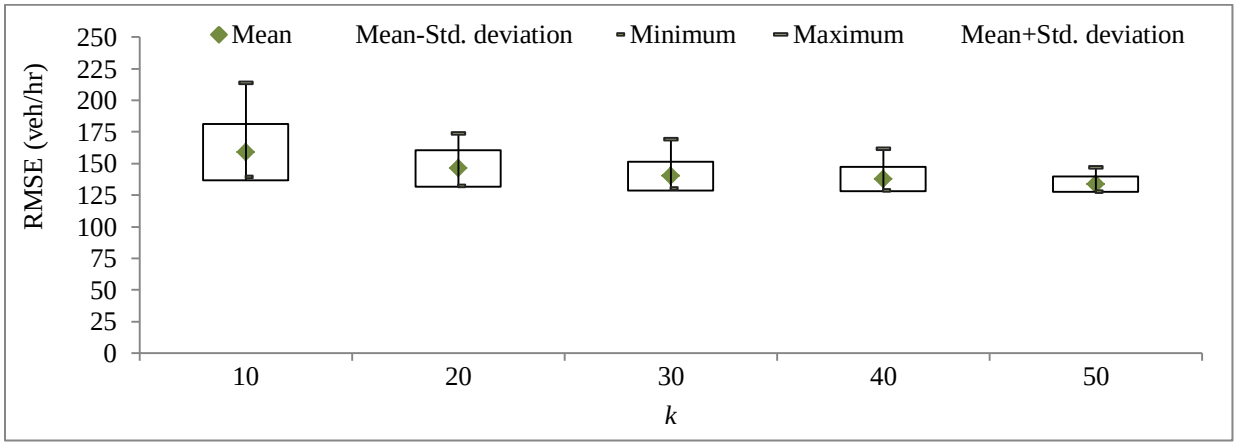


a) Mean

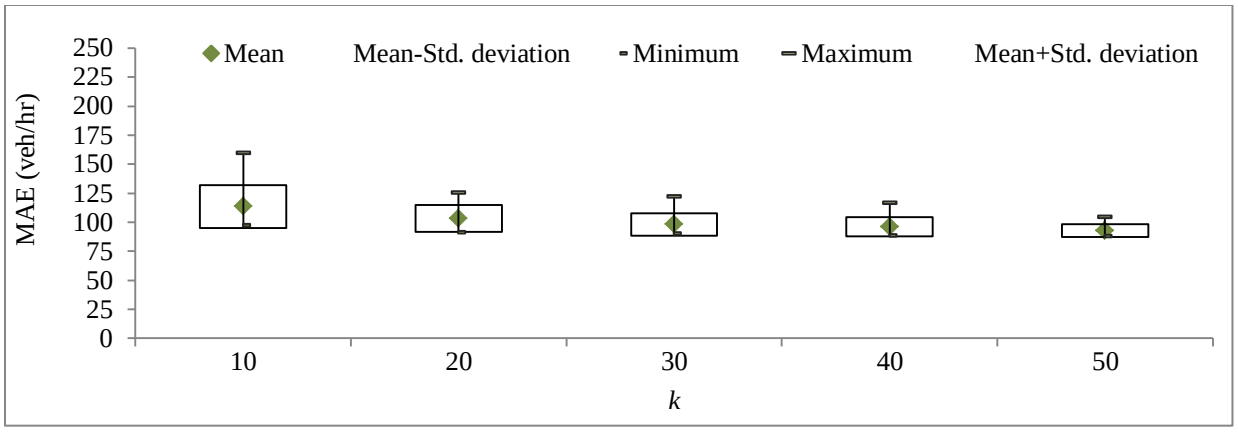


b) 90th percentile

Figure 3.5: Prediction accuracy (MAE) of short-term prediction models with top-*k* important predictors based on the four predictor selection scenarios.



a) RMSE



b) MAE

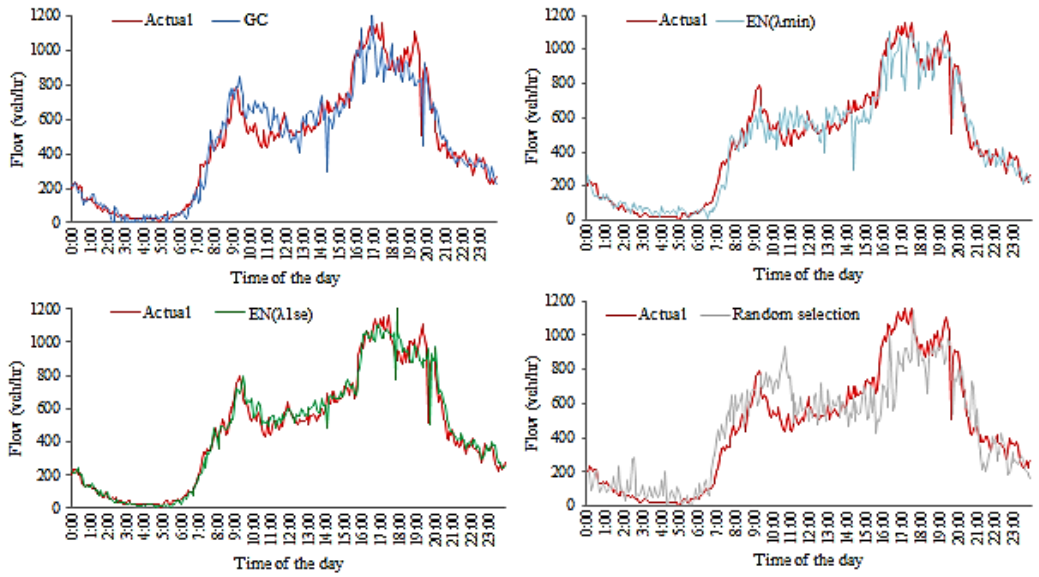
Figure 3.6: Box-plot showing the distributions of RMSE and MAE across 10 sets of randomly selected *k* predictor links in Scenario 4.

To illustrate the prediction accuracy of the four scenarios, the actual traffic flow on a selected target link is compared with the predicted flow estimated by the four scenarios under different *k* values. Figure 3.7 represents the differences of the actual and predicted flow on target link 360 for a single day. Among all the scenarios, the predicted flow based on EN(λ_{1se}) is the closest to the actual flow in all of the five *k* values (top-10, 20, 30, 40, and 50 predictors). The predicted flow under GC shows less deviation from the actual flow than the predicted flow under EN(λ_{min}) and Random. The Random scenario shows significant fluctuations in predicted flow compared to the actual traffic flow. The predicted flow of each scenario becomes closer to the actual flow as *k* (the number of predictors) increases. Overall, the results in Figure 3.7 are consistent with the results discussed with Figures 3.5 and 3.6.

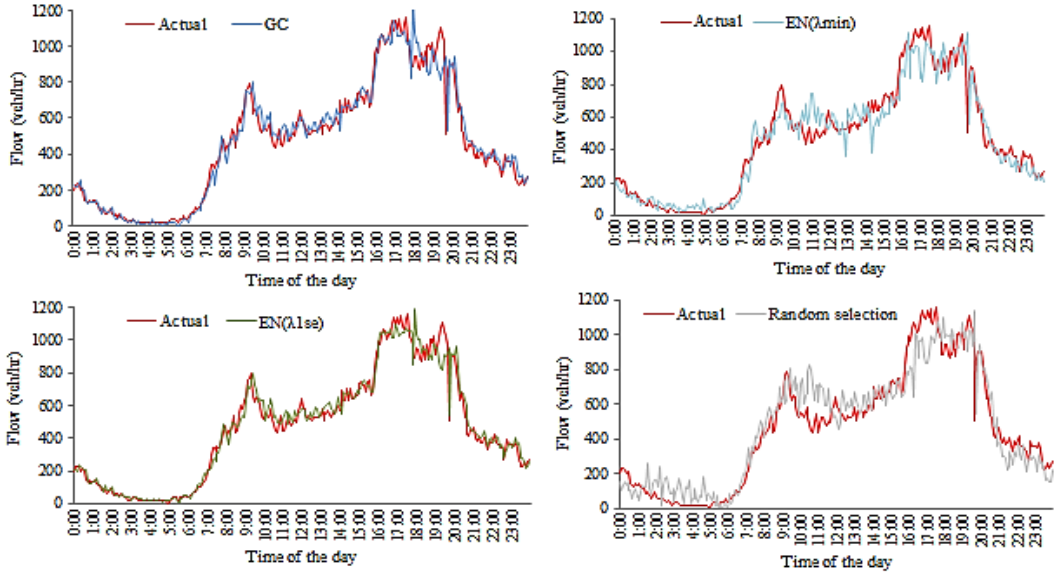
From the fact that the predicted flow of EN(λ_{1se}) reflects the actual traffic flow very closely, while achieving the largest dimensionality reduction among GC, EN(λ_{min}), and EN(λ_{1se}) (as shown in Figure 3.1), the use of EN(λ_{1se}) regularisation as a variable selection technique is found to offer the most effective way of identifying important predictors for network-wide traffic prediction in this study.

This study provides evidence that considering important variables is useful in improving the prediction accuracy of short-term traffic prediction models, but the set of important variables

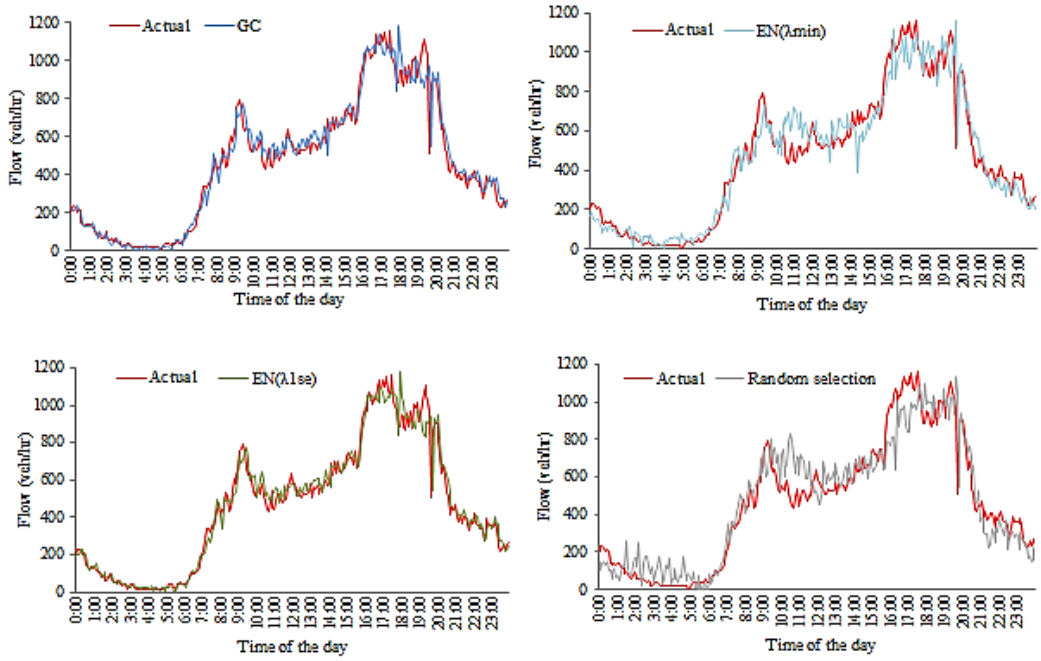
should be selected in such a way that allows to capture correctly the underlying spatial relationship because the failure to exclude irrelevant variables can decrease the model performance. The proposed methods, thus, provide a useful tool for automatically detecting the important predictors in the road network, which seem to matter more than the physical connectivity when it comes to improving the prediction accuracy, and suggesting a set of predictors for any given target link of interest.



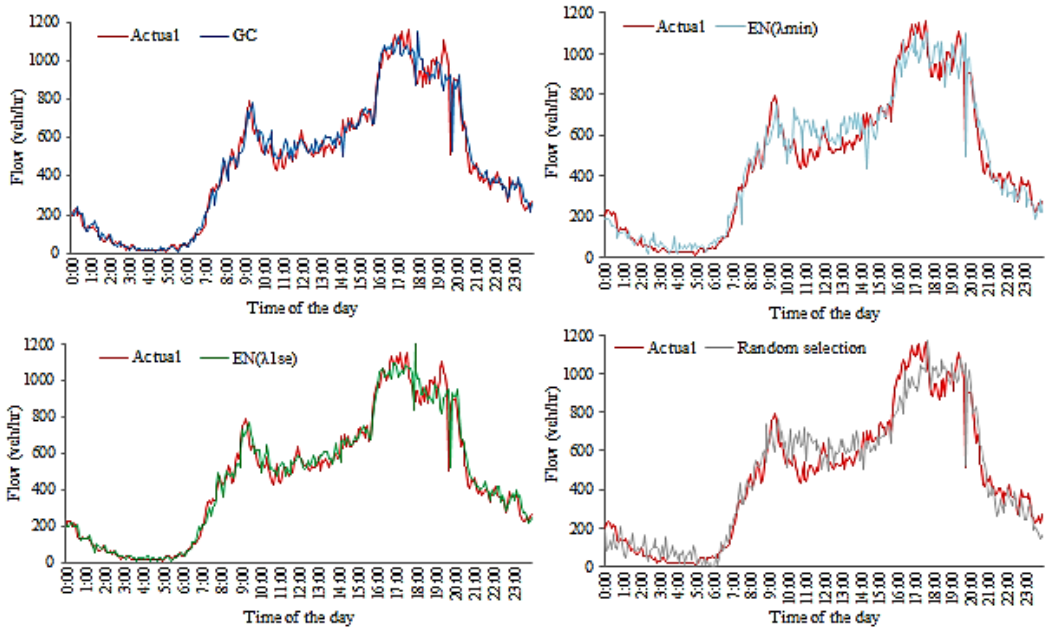
a) The 10 most important predictors



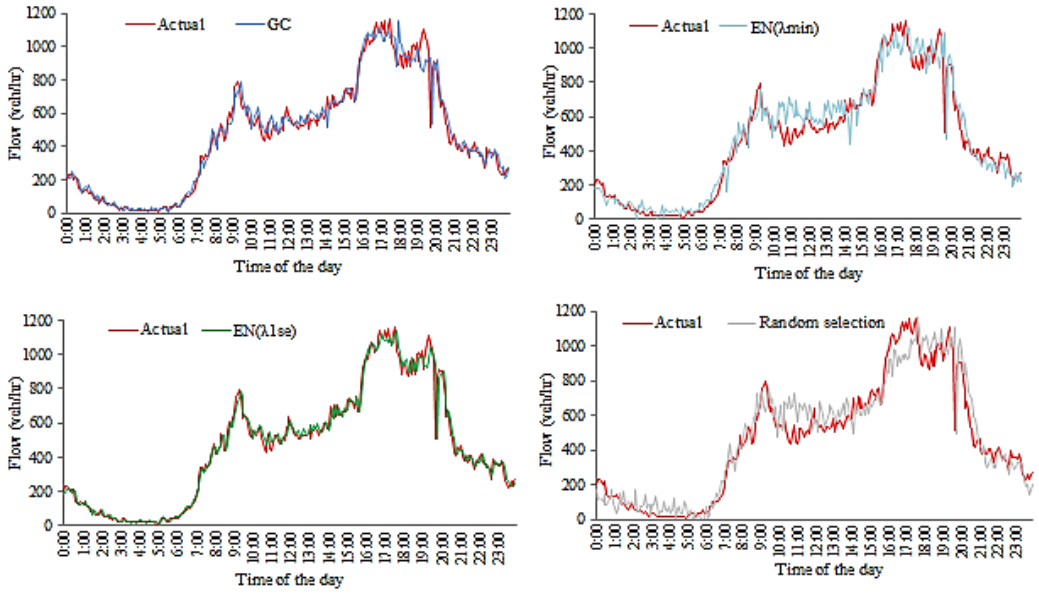
b) The 20 most important predictors



c) The 30 most important predictors



d) The 40 most important predictors



e) The 50 most important predictors

Figure 3.7: Actual and Predicted traffic flow of a single day by different set of predictor links for a target link (target link 360 as an example).

3.5 Conclusion

This study demonstrates the advantages of accounting for importance of road links in the network through considering spatial relations. The proposed method determines the importance of each road link as a predictor and identifies a hierarchy or rank of important predictor links in the road network. Therefore, a common set of most important road links in the network is obtained by taking only the higher ranked links in order to reduce the number of predictors as well as to improve the short-term traffic prediction. Short-term traffic prediction models developed by using the common set of important predictors show higher accuracy than the prediction models based on the randomly selected set of predictors. This indicates the efficacy of selecting sets of important road links as predictors rather than considering any set of road links as predictors in a short-term traffic prediction model. Also, the consistency of the prediction results shows the advantages of considering a hierarchy of importance of road links when predicting traffic states in large-scale road network. The important predictor sets identified using EN(λ_{1se}) regularisation were found to be more effective than the important predictor sets determined by GC analysis in predicting traffic flow of all the road links in the network. This suggests the

superiority of EN(λ_{1se}) in determining the hierarchy or ranking of important predictor links.

The findings from this study suggest that traffic authorities have the possibility to monitor and predict traffic states with a limited expenditure in the number of sensors. Moreover, the findings about the importance of the links in the network suggest that, in the case of malfunctioning sensors in the target location, traffic authorities may consider the location of important predictors to measure or predict the traffic parameters on the entire road network. Further research could address some of the limitations of this study. Firstly, the analysis could discern between peak and off-peak hours as this study focused on traffic states of road links over the whole day period for the implementation of the proposed method of identifying important road links in the network. Furthermore, the variation of the tuning parameter alpha in the EN regularisation could be explored for values other than the assumed 0.5, thus allowing a variation between the ridge and the lasso extremes.

Chapter 4

DETERMINING NONLINEAR SPATIAL DEPENDENCY STRUCTURE IN URBAN ROAD NETWORKS BY REGRESSION TREE METHOD

4.1 Introduction

The availability of ample data resources and fast advancement of computational power leads to growing interests in data-driven or machine learning approaches for traffic prediction in a large-scale urban road network. One of the impediments in traffic prediction is the efficient utilisation of substantial amount of data to identify a set of predictor variables that are pertinent to the target or response variable. Therefore, understanding of dependency structure among the variables in an extensive data set is beneficial for traffic prediction.

The traffic states of a given target link depend on the previous traffic states of its own and the other links in the network. If this dependency is properly captured, the future traffic states of the target link can be accurately estimated. The dependence of traffic states of a target link on the other links in the network can be represented as a linear or nonlinear relationship. Most of the existing literatures assume that the dependence relationship between the target and predictor links is linear (Stathopoulos and Karlaftis, 2003; Kamariankis and Prastacos, 2004; Sun et al., 2006; Chandra and Al-Deek, 2009). However, this relationship can be nonlinear as traffic flow is known to follow a nonlinear process (Smith et al., 2002).

A number of studies in other research areas (e.g. economics) compared the dependence of the target variables on other variables that was identified by the linear and nonlinear methods of Granger causality. They observed that the causal dependence relies on the type (linear or nonlinear) of method used. A target variable can have independence on a predictor variable in the linear causality method whereas a causal dependence can be observed in the nonlinear causality method (Pavlidis et al., 2015; Chu et al., 2016; Yu et al., 2015). Thus, it is useful to

employ a nonlinear method to detect dependence between variables if the system itself is nonlinear.

Some of the existing methods used in traffic prediction can capture the nonlinear dependence between the target and the predictor variables. These methods are typically based on non-parametric approaches such as neural network and Bayesian network. It is observed that these approaches are effective in traffic prediction when a limited number of predictor variables are available (Goves et al., 2015; Chen et al., 2011; Kumar et al., 2013; Gosh et al., 2007; Li et al., 2019; Wang et al., 2014; Pascale and Nicoli; 2011; Sun and Zhang, 2006; Queen and Albers, 2009). Using hundreds of input variables considerably increase the computation time of the approaches and restrict its capability to produce accurate results. Unless a reduced number of predictor variables are provided as the input variables, these existing nonlinear methods are not effective in traffic forecasting for a large-scale road network that includes hundreds of road links.

It is also noted that most of previous studies did not aim to capturing actual dependence among the traffic states of road links in the network. They either assumed linear dependence or considered nonlinear dependence between the traffic states of the predictor links and the target link in traffic prediction. Another research gap exists in identifying the relevant predictor links for a given link by considering actual dependence among the traffic states of the road links in a large scale road network. Therefore, an approach to identifying a parsimonious set of relevant predictors for the target link by capturing actual dependence among the road links in the road network is required to ensure an accurate traffic prediction within a short computation time.

Given the need to find a systematic way of identifying a set of relevant links for the given target road link by capturing the actual (linear or nonlinear) dependence among road links in the network, this study proposes a statistical approach that uses the regression tree method (Breiman et al., 1984; Therneau and Atkinson, 1997). The proposed approach primarily considers all road links in the network as possible predictors of a given target link and then selects the relevant ones by the regression tree. The regression tree does not assume linear dependency of traffic parameter of the target and the predictor links but explores the underlying relationship in a non-parametric way, allowing to capture complex nonlinear relationships in the data. Regression tree

is a data mining method and unlike other 'black box' type data mining methods (e.g. neural network), this method is comparatively easier to comprehend. Regression tree has been used in various research areas including traffic prediction (Hou et al., 2014) and road safety (Karlaftis and Golias, 2002; Siddiqui et al., 2012). However, regression tree was not employed in previous studies for identifying the relevant predictor links of a given link by using spatial dependency of the road links in a large-scale urban road network. Our proposed method is intended to serve as an input variable selection method for a traffic prediction model, where a set of road links should be selected as model input such that all the input predictor links are informative in predicting the traffic state of the target link. Accordingly, the method helps improve the prediction accuracy while keeping the number of variables to a minimum to obtain a parsimonious model. It should be noted that the focus of this study is on the development of an 'input variable selection method' for a short-term traffic prediction model, not the traffic prediction model itself.

This study also compares the effectiveness of the proposed nonlinear method with a linear method of relevant predictor selection in terms of prediction accuracy and model dimensionality reduction. The linear method applied for the comparison uses the Granger causality analysis (Granger, 1969; 1980), which assumes linear relationship of traffic flow between the target link and the predictor links. The Granger causality test is proven to be efficient in selecting the relevant predictor links for the given target link in the network (Hasan and Kim, 2016; Hasan et al., 2017).

The remainder of the chapter is organised as follows. Section 4.2 describes the nonlinear approach of spatial variable selection, its implementation and performance evaluation. Section 4.3 illustrates the case study and details the application of the proposed method to the large-scale road network. Section 4.4 illustrates and discusses the results of the case study before section 4.5 presents the conclusion of this study and proposes future research avenues.

4.2 Selection of relevant predictor links by nonlinear and linear methods

4.2.1 Regression tree as the nonlinear method of predictor selection

Regression tree is a type of decision tree whose predictor variables and target variable take continuous values. This is a nonparametric approach which explores the structure in the dataset without assuming an underlying distribution. Regression tree can handle the nonlinear and complex interactions of numerous features within the dataset. This method is simple and easy to interpret (Lewis, 2000). Regression tree can be developed by recursive partitioning technique (Therneau and Atkinson, 1997). Recursive partitioning technique segregates data into smaller, non-overlapping and homogenous subsets to detect the interactions of the target and predictor variables in a more manageable way. The procedure is repeated over each of the smaller division until it becomes infeasible to continue. The divisions of data made by recursive partitioning technique are represented by the nodes in a tree-like structure.

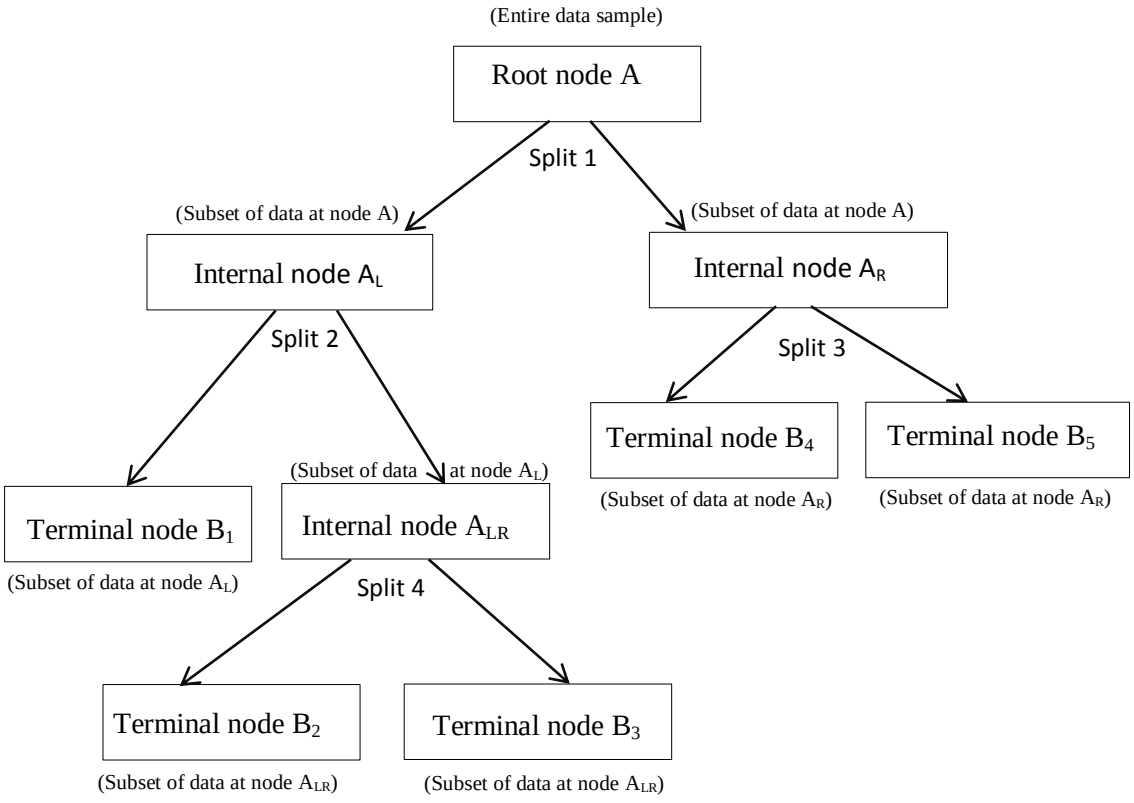


Figure 4.1: Graphical representation of a regression tree structure.

Regression tree has three main parts: root node, internal node and terminal node. The nodes in the regression tree are developed by using splitting process. Splitting is a procedure of partitioning the data sample into two parts or dividing a node into two child nodes by using a threshold value of a predictor variable. The root node is the starting point of the tree which represents entire data sample whereas other nodes in the tree represent a subset of the entire data sample. Using the splitting process, the root node produces two child nodes which, in turn, become parent nodes and further produce additional child nodes. The procedure continues until it reaches to the terminal nodes (i.e. end of developing new nodes in the tree). A fully grown tree along with a number of terminal nodes are obtained when all observations in the child nodes have same distribution of the parent node or only one observation exists in each child node (Fonarow et al., 2005). The nodes between the root node and terminal nodes are called internal nodes. The value in each terminal node represents an output of the regression tree which is the mean value of all available observations of the target variable in a pathway from the root node to that terminal node. Therefore, each terminal node in the regression tree possesses different value. The fully grown tree developed by using recursive partitioning technique can experience over-fitting (Lewis, 2000; Lawrence and Wright, 2001). To minimise the effect of over-fitting and reduce the complexity of the tree, the fully grown regression tree can be pruned.

For this study, consider a road network with *N* number of road links. Among the road links, the traffic states of the target link *n* is to be estimated by utilising the past traffic states of the all the available road links *N* as the predictor links. Let the future traffic states of the target link is $\theta_{n,t}$ and the past traffic states of the predictor links $\theta_{1,t-p}, \theta_{2,t-p}, \cdots, \theta_{N,t-p}$ with time lag $p = 1, 2, \cdots, P$. The focus of the study is to identify the past traffic states of predictor links that are more relevant to estimate the traffic states of the target link. This can be achieved by employing the regression tree as the variable selection method. The regression tree divides the traffic states data of the target link ($\theta_{n,t}$) into several smaller subsets, each of which is associated with a different range of the traffic states of the predictor links ($\theta_{1,t-p}, \theta_{2,t-p}, \cdots, \theta_{N,t-p}$). The partitions of data (or building of nodes) brings more homogenous clusters of the target link states together which aids the accurate estimation of the traffic states of target link ($\theta_{n,t}$). Therefore, identification of the road links whose traffic states made the partition of data, can be an approach in selecting spatial variable link in a road network.

In the regression tree, initially, $\theta_{1,t-p}, \theta_{2,t-p}, \cdots, \theta_{N,t-p}$ are taken as the predictors and $\theta_{n,t}$ is considered as the target variable. All the traffic states data of the target link ($\theta_{n,t}$) are grouped into the same partition which is exemplified by the root node *A* of the regression tree in Figure 4.1. In the next step, the data of $\theta_{n,t}$ are divided into two subsets which are represented as the two child nodes (A_L and A_R) of root node *A*. These two subsets of data or two child nodes are obtained after comparing every possible split using every traffic state value of each of the predictors $\theta_{1,t-p}, \theta_{2,t-p}, \cdots, \theta_{N,t-p}$. For instance, given a predictor θ_x in the predictor set, i.e., $\theta_x \in \{\theta_{1,t-p}, \theta_{2,t-p}, \cdots, \theta_{N,t-p}\}$, and a threshold value of h, the target link data in the root node (A) can be divided into two subsets: one (A_L) containing the data of target link $\theta_{n,t}$ where predictor $\theta_x < h$ and the other (A_R) containing the data of target link $\theta_{n,t}$ where predictor $\theta_x \geq h$.

Among all the possible combinations of predictor θ_x and threshold h, the splitting condition ($\theta_x < h$ and $\theta_x \geq h$) that offers the "optimal split" for the division of the target link data—the predictor condition that leads to two of the most homogenous subsets of target link states—is considered as the threshold value of the relevant predictor for this partition of data. The optimal split can be identified by using residual sum of squares (RSS). For instance, if the root node *A* is to be split into two child nodes A_L and A_R, it requires to satisfy the following condition (Therneau and Atkinson, 1997):

$$RSS(A_L) + RSS(A_R) < RSS(A) \tag{4.1}$$

The split ($\theta_x < h$ and $\theta_x \geq h$) which maximises the difference between the right-hand side and the left-hand side of Eq. (4.1), denoted by Δ, is selected as the optimal split as follows:

$$\max_{\theta_x,h} \Delta = \max_{\theta_x,h} RSS(A) - \{RSS(A_L) + RSS(A_R)\} \tag{4.2}$$

where, $RSS(A)$ is the residual sum of squares for root node *A*; $RSS(A_L)$ and $RSS(A_R)$ are the residual sum of squares for root node *A*'s child node A_L and child node A_R, respectively. The residual sum of squares of the root node and its child nodes are estimated by the following equations:

$$RSS(A) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y}_A)^2 \quad (4.3)$$

$$RSS(A_L) = \frac{1}{T1}\sum_{t1=1}^{T1}(y_{t1} - \bar{y}_{AL})^2 \quad (4.4)$$

$$RSS(A_R) = \frac{1}{T2}\sum_{t2=1}^{T2}(y_{t2} - \bar{y}_{AR})^2 \quad (4.5)$$

where, y_t is the value of the target variable at observation *t* in the set of data for node *A* and $\bar{y}_A$ is the mean value of the target variable in the subset of data for node *A*. Similarly, y_{t1} and y_{t2} are the values of target variable at observation t_1 and t_2 in the subset of data for child nodes A_L and A_R, respectively; $\bar{y}_{AL}$ and $\bar{y}_{AR}$ are the mean values of the target variable in the subset of data for child nodes A_L and A_R, respectively. The size of data for node *A*, denoted by T, is the sum of the size of data for A_L, denoted by $T1$, and the size of data for A_R, denoted by $T2$ (Therneau and Atkinson, 1997).

Similarly, other internal nodes are developed and the input traffic states data are divided into a number of partitions by the splitting procedure. In each internal node, the threshold traffic state value of a predictor variable among the predictors (i.e. $\theta_{1,t-p}, \theta_{2,t-p}, \cdots, \theta_{N,t-p}$) is obtained. The splitting procedure continues until it becomes infeasible to partition the data and thus all the terminal nodes in the tree (e.g. nodes B_1-B_5 in Figure 4.1) are developed. Each of the terminal nodes is an output of the regression tree which displays the mean value of all available observations of the target variable in a pathway from the root node to that terminal node mean of the subset of $\theta_{n,t}$.

Pruning procedure reduces the nodes of the fully grown regression tree, and the pruned tree performs better than fully grown tree in the context of new data cases (Therneau and Atkinson, 1997). In this study, we adopt a post pruning method by using an advisory parameter known as cost complexity parameter (α). This parameter specifies how a tree is penalised by additional splits. The value of α is estimated by the following equation:

$$R_\alpha(S) \equiv R(S) + \alpha * |S| * R(S_0) \quad (4.6)$$

where T_0 is the tree having no split, $|S|$ is the number of splits for a tree, and R is the risk which can be considered as the cross-validation error, $R(S_0)$ is the risk for zero split tree, $R(S)$ is the risk of a tree having S splits and $R_\alpha(S)$ represents regularised risk of the tree having S splits (Therneau and Atkinson, 1997). The full regression tree is built first and the tree is then pruned using optimal α value. Higher α value results over-pruning of the tree and lower α value selects larger tree. The one with least cross-validated error is considered as the optimal value of α. The cross-validation error of each split of the tree can be obtained by using *n*-fold cross-validation. In *n*-fold cross-validation, the data are divided into *n* randomly selected sets. Each of the *n* sets of the data serves as a validation set once and serves as the training sets for remaining *n*-1 times. Every split from the root node to the terminal node is fitted by using each *n*-1 set of data and the corresponding cross-validation error is estimated. The average cross-validation error for *n* folds is computed. The split, where the minimum value of the average cross validation error was obtained, is considered as the terminal split and the corresponding α value is noted as the optimal α value. The nodes of the fully grown tree after the terminal split are pruned by using the optimal α value since no further improvement in cross-validation error is obtained.

The predictor links which are used in the splitting process or developing the nodes in the regression tree can be considered as the relevant predictor links for the target link. A splitting process occurs when a partition of dataset can reduce the RSS of estimating the traffic states of the target link. The partitioning of the dataset into smaller subsets helps to reduce the RSS of the estimation as smaller subsets contain homogenous data. Therefore, the predictor links whose threshold traffic state value offers the optimal split for partition of the dataset (i.e. developing two child nodes from a parent node in the regression tree), is considered to be the relevant predictor links for the target link since this predictor is more useful than other predictors in estimating the traffic states of target variable. However, the predictors which are not used in building the tree cannot be entirely considered as irrelevant to the target variable since some of them can also be used as the surrogate predictors when missing data are encountered. If the traffic state of a predictor link is selected more than once for splitting the dataset in a single regression tree, this predictor link is also considered as a single relevant predictor link.

In this study, each link in the road network is considered as a target link once and taken as the predictors for other target links. For each target link, a distinct regression tree can be obtained. The nodes of each regression tree are developed by the traffic states of a number of predictor links. The predictor links which traffic states are used in building the regression tree for a target link are taken as the relevant predictor link for the target link. These relevant predictor links can be useful in the prediction model to forecast traffic states of the target link.

4.2.2 Granger causality as the linear method of predictor selection

The *Granger causality* analysis recognises directed functional or causal interactions of different variables in time series data and excludes variables without an interaction (Seth et al., 2015; Li et al., 2015). The Granger causality analysis measures the effect of the histories of a variable $(\text{e.g.}\, x_{t-1})$ in predicting another variable $(\text{e.g.}\, y_t)$ by comparing the prediction errors. If inclusion of x_{t-1} reduces the prediction error significantly, then it is considered to be an improvement of the prediction of y_t due to the inclusion of x_{t-1} and it is said that time series x "Granger causes" time series y, according to Granger causality (Granger, 1969). In the traffic analysis context, the Granger causality analysis allows to find out the spatial relations between road links in a road network. Unravelling these relations is then used to identify relevant road links that can be used to predict the traffic parameters of a given target road link.

This study adopts VAR based Granger causality to identify relevant predictor link for a given target link in the road network. In a VAR model, each variable is taken as response variable once and the other variables are considered as explanatory variables. Multivariate time series analysis is used in a VAR model where each variable is computed by its own lagged values and the lagged values of other variables (Zivot and Wang, 2006). Let $\boldsymbol{\theta}_t = (\theta_{1,t}, \theta_{2,t}, \cdots, \theta_{N,t})$ be an $(N \times 1)$ vector representing a traffic flow measure for *N* road links at time *t*, where $t = 1, 2, \ldots T$. The basic *P*-lag VAR model for predicting $\boldsymbol{\theta}_t$ can be written as follows:

$$\boldsymbol{\theta}_t = \mathbf{c} + \boldsymbol{\Pi}^{(1)}\boldsymbol{\theta}_{t-1} + \boldsymbol{\Pi}^{(2)}\boldsymbol{\theta}_{t-2} + \cdots + \boldsymbol{\Pi}^{(p)}\boldsymbol{\theta}_{t-p} + \cdots + \boldsymbol{\Pi}^{(P)}\boldsymbol{\theta}_{t-P} + \boldsymbol{\varepsilon}_t \tag{4.7}$$

where $\mathbf{c}$ is an $(N \times 1)$ vector of intercepts and $\boldsymbol{\Pi}^{(p)}$ is an $(N \times N)$ coefficient matrix reflecting the relationships between the response variable $(\boldsymbol{\theta}_t)$ and its *p*-lagged variable $(\boldsymbol{\theta}_{t-p})$ with time

lag $p = 1, 2, \cdots, P$. Eq. (4.7) can be decomposed to show the components related to a single target link n ($n = 1, 2, \cdots, N$) as follows:

$$\theta_{n,t} = c_n + \left(\pi_{n,1}^{(1)}\,\theta_{1,t-1} + \pi_{n,2}^{(1)}\,\theta_{2,t-1} + \cdots + \pi_{n,N}^{(1)}\,\theta_{N,t-1}\right) + \left(\pi_{n,1}^{(2)}\,\theta_{1,t-2} + \pi_{n,2}^{(2)}\,\theta_{2,t-2} + \cdots + \pi_{n,N}^{(2)}\,\theta_{N,t-2}\right) + \cdots + \left(\pi_{n,1}^{(P)}\,\theta_{1,t-P} + \pi_{n,2}^{(P)}\,\theta_{2,t-P} + \cdots + \pi_{n,N}^{(P)}\,\theta_{N,t-P}\right) + \varepsilon_{n,t} \quad (4.8)$$

where $\theta_{n,t}$ denotes a scalar at the n^{th} element of $\boldsymbol{\theta}_t$ representing the traffic flow measure for link n at time t and $\pi_{n,m}^{(p)}$ is a scalar at the n^{th} row and the m^{th} column of $\boldsymbol{\Pi}^{(p)}$ representing the relationship between the traffic condition on link n at time t (target link or response variable) and the traffic condition on link m at time $t - p$ (predictor link or explanatory variable), where $n, m = 1, 2, \cdots, N$.

To determine the time lag order P, model selection criteria such as the Akaike Information Criterion (AIC) and the Schwarz-Bayesian Information Criterion (BIC) can be used. At first the VAR model is fitted with each of the lag orders $P = 0, 1, \cdots, P_{max}$ and the corresponding value of the model selection criterion is calculated. Then, the actual time lag order can be identified by comparing the scores of the AIC and BIC:

$$\text{AIC} = -2\ln(L) + 2k \quad (4.9)$$

$$\text{BIC} = -2\ln(L) + k\ln(T) \quad (4.10)$$

where L is the maximised value of likelihood function of the model at the value of the parameter estimates, k is the number of parameters in the model, and T is the number of observations. It should be noted that the BIC or AIC scores decrease with an improvement in the log-likelihood and a decrease in the number of parameters. The lag order with the lowest AIC or BIC score is considered as the best lag order for modelling (Cottrell and Lucchetti, 2016). Since one of the objectives of this study is to reduce the number of parameters in the model and obtain a parsimonious model, the lowest lag order between AIC and BIC is selected in this study.

After the specification of the VAR model, Granger causality between a target link and each of its predictor links is tested by performing a F-test. Consider target link n and predictor link m. A time series of predictor link $\theta_{m,t}$ is considered to Granger-cause a time series of target link $\theta_{n,t}$ if at least one of the lagged values of $\theta_{m,t}$ provides stastically significant information about future values of $\theta_{n,t}$. This can be tested through the F-test with the null hypothesis $\mathrm{H}_0: \pi_{n,m}^{(1)} = \pi_{n,m}^{(2)} = \cdots = \pi_{n,m}^{(P)} = 0$ and the alternative hypothesis $\mathrm{H}_1: \left(\pi_{n,m}^{(1)} \neq 0\right) \cup \left(\pi_{n,m}^{(2)} \neq 0\right) \cup \cdots \cup \left(\pi_{n,m}^{(P)} \neq 0\right)$. The null hypothesis that $\theta_{m,t}$ does not Granger-cause $\theta_{n,t}$ is rejected if at least one of the elements $\pi_{n,m}^{(p)}$ for $p = 1, 2, \cdots, P$ is significantly larger than zero (Bahadori and Liu, 2012). The F-test statistic is computed as follows:

$$F_0 = \frac{\frac{RSS_r - RSS_{ur}}{v}}{\frac{RSS_{ur}}{T-(q+1)}} \tag{4.11}$$

where RSS_r is the sum of the squared residuals of a restricted model (e.g., the model with $\pi_{n,m}^{(1)} = \pi_{n,m}^{(2)} = \cdots = \pi_{n,m}^{(P)} = 0$), RSS_{ur} is the sum of the squared residuals of an unrestricted model (e.g., the full model in Eq. (4.7)), v is the number of restrictions or the number of coefficients being jointly tested, T is the number of observations, and q is the number of explanatory variables in the unrestricted model. The F_0 value is then compared with the critical value of F at the 0.01 significance level. If the F_0 value is higher than the critical value, it rejects the null hypothesis which means the rejection of the statement that the time series of traffic state on the tested predictor link does not Granger-causes the time series of traffic state on the target link. Otherwise, the null hypothesis cannot be rejected indicating that the tested predictor link does not provide statistically significant information about future states on the target link.

4.3 Case study

For this case study, the road network and the traffic state data as described in section 2.3.1 and section 2.3.2 are nominated. The traffic state data of each of the 479 road links in the road network are considered as a target link once and the traffic states of the remaining road links are selected as predictor links.

4.3.1 Nonlinearity test

Before applying the proposed nonlinear method, the likelihood of traffic states of the target link and predictor link having a linear or nonlinear relationship has been analysed. This study adopts a certain kind of misspecification test named as Ramsey Regression equation specification error test.

4.3.1.1 Ramsey Regression equation specification error test (RESET test)

RESET test (Ramsey 1969; 1974) is a misspecification test where a linear functional form between the target and the predictor variable is compared with a nonlinear (e.g. polynomial) functional form. Let a linear regression between the target link y_t and the past of the predictor link x_{t-1} be as follows:

$$y_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon_t \tag{4.12}$$

where β_0 and β_1 are the intercept and the coefficient, respectively, and ε_t is the prediction error. The RESET test explores whether a nonlinear function $(\beta_1 x_{t-1})^2, (\beta_1 x_{t-1})^3, \dots, (\beta_1 x_{t-1})^k$ can have any significance in estimating y_t. This can be written as

$$y_t = \beta_0 + \beta_1 x_{t-1} + \gamma_1 (\hat{y}_t)^2 + \gamma_2 (\hat{y}_t)^3 + \cdots + \gamma_k (\hat{y}_t)^k + \varepsilon_t \tag{4.13}$$

Then, the F-test is employed to determine whether the coefficients of the polynomial coefficients $\gamma_1, \gamma_{2,\dots,}\gamma_k$ are zero or nonzero. The null hypothesis is that the linear functional form of the regression between the target and the predictor variable is sufficient to estimate y_t (i.e. $\gamma_1 = \gamma_2 = \gamma_k = 0$). If the null hypothesis is rejected, then the model is susceptible to misspecification. This means that a nonlinear functional form should be used in the regression.

For randomly selected 15 target links, we observed that the null hypothesis is rejected at a significance level of 0.05, which means that the nonlinear functional relationship prevails between the target link and the predictor link.

The following figures show the relationship of traffic flow between target link and the predictor links considering stationary data (Figure 4.2). For instance, Link 80 is taken here as the target link and three road links (Link 79, Link 236 and Link 234) are considered as the predictor links to show the relationship graphically. As we used de-trended data, a number of negative values of traffic flow can be seen in Figure 4.2. The X axis represents the de-trended traffic flow of the target link and the Y axis represents the de-trended traffic flow of the predictor link. It can be observed that traffic flow of the target link and the predictor links are not in a linear relationship. Also, the relation between the target link and each of the predictors varies.

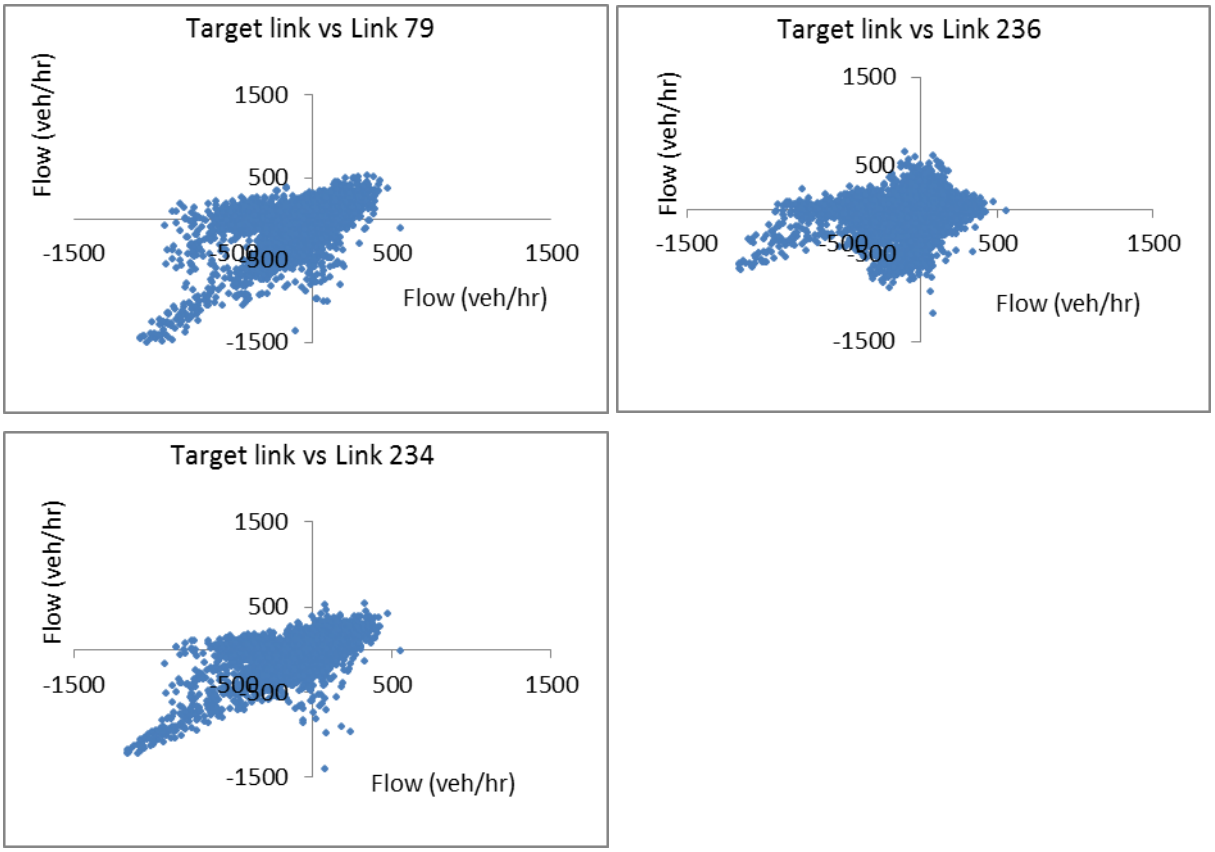


Figure 4.2: Relationship of traffic flow of the target link (Link 80 as an example) and predictor links based on stationary data.

4.3.2 Development of the nonlinear and linear method of relevant predictor selection

4.3.2.1 Building Regression tree

The selected network consists of 479 road links with available traffic data. In this case study, we develop a regression tree model by considering each of the road link as the target link once and as the predictor for the remaining target links. In predicting each target link, the past traffic flow of the target link and 478 links are considered as the initial predictors. Lag order1 is selected based on the analysis results of BIC (Eq. (4.10)). Separate regression tree is built for each of the target links. The full regression trees are built by using the splitting rules that were mentioned in section 4.2.1. The growth of the tree is terminated by the stopping criterion, which states that the minimum number of observations in a terminal node is one and the minimum number of observations in a node for splitting is two. The fully grown regression trees are then pruned by using cross-validation error and cost complexity parameter. 10-folds cross-validation process is implemented in the entire training data to find the cross-validation error for each split in the full regression tree. The optimal cost complexity parameter value, at which the minimum cross-validation error was found, is used to prune the full tree. All the nodes of the regression tree, whose corresponding cost complexity parameter value is higher than the optimal cost complexity parameter value, are retained in the pruned tree. Pruning of the regression tree delivers a tree with a reduced number of nodes. The predictor links, whose traffic flows are used to create the nodes in the regression tree for a target link, are considered as the relevant predictor links for that target link.

In this study, the regression tree is implemented using R software package called rpart (Therneau and Atkinson, 1997). The regression tree is developed for each of the 479 target links in the road network. Each target link has a distinct regression tree diagram, where different predictors are used to develop the branches or nodes of the tree. These predictors are considered as the relevant predictors for the target link. Figure 4.3 illustrates an example of a regression tree diagram of a target link titled as Link 80.

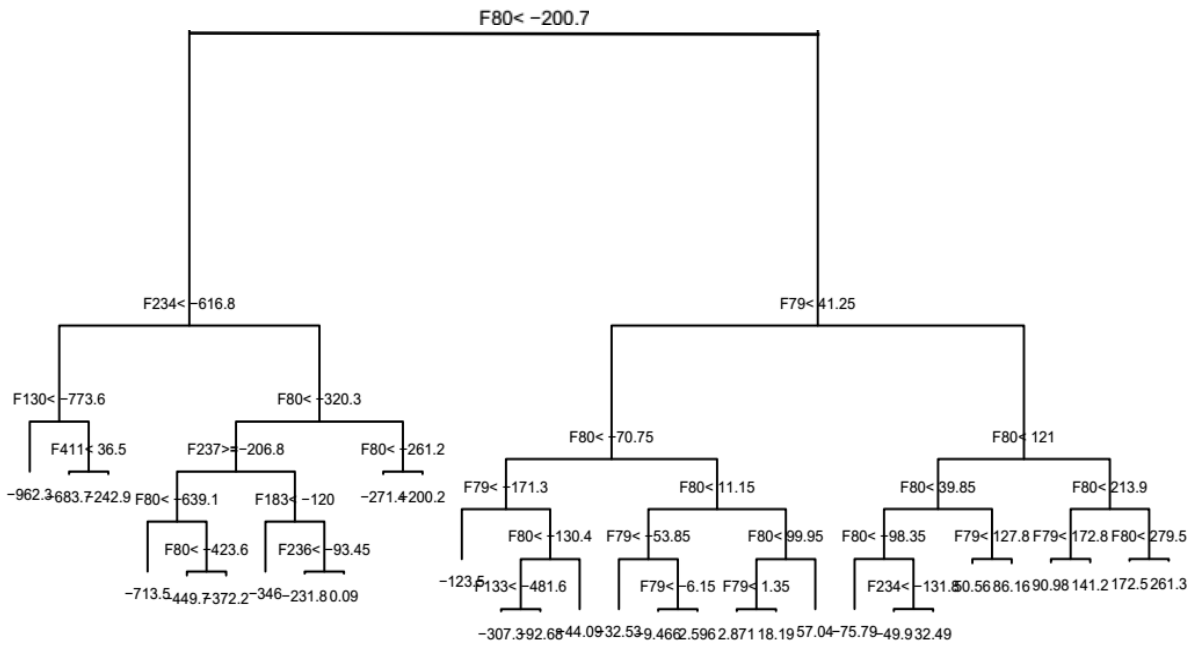


Figure 4.3: Graphical representation of the regression tree for a target link (Link 80 as an example).

In Figure 4.3, F80 represents one lagged past traffic flow of Link 80 and the corresponding value represents the flow value which has unit veh/hr. As we used de-trended data, we can see some negative values of traffic flow in the nodes of the regression tree. The root node was created by the past traffic flow of the target link (Link 80) in this example; however, it may not be the case for the regression tree of other target link. The whole traffic flow dataset for building regression tree is divided into two parts by a threshold traffic flow value of F80. The threshold value F80 <-200.7 in the root node actually represents a question, 'Is F80 is less than -200.7 veh/hr?' If it is true, then the subset of whole data corresponding to F80<-200.7 veh/hr belongs the left side of the root node and if it is false, then the data corresponding to F80 >-200.7 veh/hr belongs the right side of the root node. Then the subset data belong to the left side of the tree are further partitioned by the threshold value of one lagged past traffic flow of link 234 (i.e. F234<-616.8 veh/hr). Similarly, the subset data belong to the right side of the tree are further partitioned by the threshold value of past one lagged traffic flow of link 79 (F79<41.25 veh/hr). These two subsets of data are subdivided into several parts until it reaches to the terminal node. The terminal nodes provide the average value of the next future interval of the target link (Link 80).

For instance, the left-most terminal node of this regression tree shows a de-trended traffic flow value as -962.3 veh/hr. This is the average value of traffic flow of the target link (Link 80) when the subset data belong to the following values: F80 less than -200.7 veh/hr, F234 less than -616.8 veh/hr and F130 less than -773.6 veh/hr. The predictor variables namely F80, F234, F130, F411, F237, F183, F236, F79, F133 are used in segregating the data into smaller partitions (i.e. developing the nodes of the regression tree) and therefore, these predictors are considered as the relevant predictors for forecasting the future traffic flow of the target link (Link 80).

4.3.2.2 Testing Granger causality

Considering each of 479 links as a target link, we apply Granger causality (GC) to identify reduced predictor links for the given target link. Lag order 1 is selected based on the analysis results of BIC. The Granger causality analysis is implemented using software package called 'Gretl' (Cottrell and Lucchetti, 2016).

4.3.3 Assessment of the efficiency of selected relevant predictors in traffic prediction

4.3.3.1 Development of short term traffic prediction

As the proposed method can capture the nonlinear relation between the target link and predictor links, a nonlinear type short term prediction model is used for the effectiveness assessment of the relevant predictors in traffic prediction. For the nonlinear type of prediction model, we adopt a multi-layer feed forward neural network with the back-propagation learning algorithm, which is known as multi-layer perceptron (Bishop, 1995). A simple multi-layer perceptron includes an input layer, one or more hidden layers, and an output layer. Each layer contains one or more neurons. The neurons are connected by directed edges which are known as synapses. The synapses are associated with a weight representing the strength of the connection between two neurons.

In this study, traffic states of relevant predictor links are used as input variables and the traffic state of the target link is taken as output variable in the neural network. The input variables include traffic states of relevant predictor links obtained from the proposed regression tree. The input variables in neural network are identical as the input variables in the multiple linear

regression model described above. One hidden layer is used for the simplicity of the analysis. The output layer has one neuron representing the response variable, which is the traffic state of the given target link. The number of neurons in the input layer varies with target link as a different number of predictor links are obtained for different target links. Thus, the total number of neurons in the hidden layers also varies. For simplicity of the prediction model, the total number of neurons in the hidden layer is selected by trial and error. The package "neuralnet" in R programming language (Günther and Fritsch, 2010) is used to build the neural network. Similar to the linear regression, the data set is divided into two parts, where 80% of data are used as a training dataset to train the neural network and the rest 20% of data are used as a testing dataset to test the accuracy of prediction. To facilitate model training, the data are normalised based on min-max method $\left(\frac{X-X_{min}}{X_{max}-X_{min}}\right)$ and are scaled in the interval of [-1, 1].

4.3.3.2 Evaluation of the prediction accuracy

This section discusses approaches to evaluating the proposed input variable selection method based on regression tree. The goal of this study is to develop an input variable selection method for a short-term traffic prediction model which can capture nonlinear and complex relation among the target link and the predictor links. As such, the performance of an input variable selection method should be evaluated based on its contribution to improving the prediction accuracy of a given traffic prediction model. The evaluation procedure consists of the following steps:

1. For a given target link, select a set of predictor links using the input variable selection method to be evaluated.
2. Build a short-term traffic prediction model using the model type of choice (e.g., non-linear neural networks), where the past traffic states of the target link and the selected predictors links from Step 1 are included as input variables. The traffic state of the target link is set to output variable.
3. Measure the prediction accuracy of the short-term prediction model and evaluate the effectiveness of the chosen input predictors in improving the prediction performance.

In short-term traffic prediction, the forecasting horizon is typically less than an hour (Smith et al., 2002; Pascale and Nicoli, 2011). In this study, the 5-minute time horizon is used and the short-term traffic prediction is implemented using multi-layer feed forward neural network.

The prediction accuracy of each short term prediction model for a particular target link m is measured using the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE) as follows:

$$RMSE_m = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\theta_{m,t} - \hat{\theta}_{m,t}\right)^2} \tag{4.14}$$

$$MAE_m = \frac{1}{T}\sum_{t=1}^{T}\left|\left(\theta_{m,t} - \hat{\theta}_{m,t}\right)\right| \tag{4.15}$$

where T is the number of observations, $\theta_{m,t}$ is the observed value of a traffic state of target link m at time t, $\hat{\theta}_{m,t}$ is the associated model predicted value.

To evaluate the overall performance of each model specification, the average of $RMSE_m$ and MAE_m over all target links $m = 1, 2, \ldots, M$ are also computed as a summary measure.

$$Mean\ RMSE = \frac{1}{M}\sum_{m=1}^{M} RMSE_m \tag{4.16}$$

$$Mean\ MAE = \frac{1}{M}\sum_{m=1}^{M} MAE_m \tag{4.17}$$

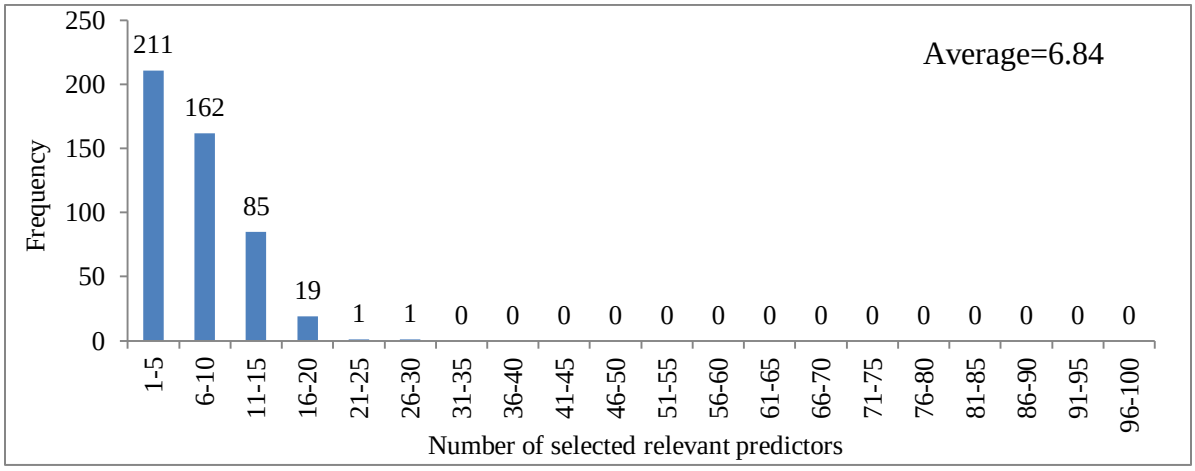
4.4 Results and discussions

This section compares the effectiveness of the proposed regression tree method over a linear method based on the vector autoregressive Granger causality (GC) test in terms of their ability to reduce the model dimensionality and enhance the prediction accuracy.

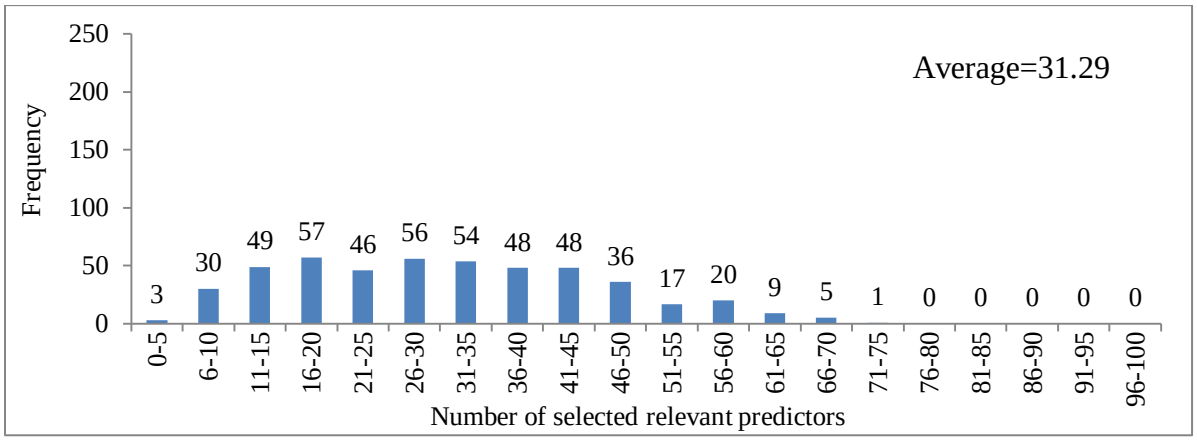
4.4.1 Dimensionality reduction

Both the regression tree and GC method have the ability to remove irrelevant predictors and select the relevant predictor links for a given target link. Each method selects a distinct set of

relevant predictors for each of the 479 target links. The detailed distributions of the number of selected relevant predictors by the regression tree across all target links are presented in Figure 4.4(a), where the range of possible values for the number of selected predictors is divided into 5 bins to create a histogram. The distribution of the number of relevant predictors selected by the regression tree is highly right-skewed. On the other hand, Figure 4.4(b) represents that the distribution of the number of relevant predictors selected by the GC method is not right-skewed but approximately a normal distribution. In comparison, these distribution diagrams suggest that the proposed regression tree generates a more parsimonious model where only the most relevant predictors are selected.



a) Regression tree



b) Granger causality method

Figure 4.4: Distribution of the number of relevant predictors selected by a) the regression tree b) the Granger causality method.

Figure 4.5 compares the average number of relevant predictors selected by the regression tree and GC. Both methods reduce a significant number of predictors and select a parsimonious set of relevant predictors. On average, the regression tree selects approximately one-fourth of the relevant predictors selected by GC. The numbers of relevant predictors identified by the regression tree and GC for each target link are illustrated in Figure 4.6. For each target link, the selection of relevant predictors by these two methods is different. Thus, the number of selected relevant predictors also varies. The regression tree method selects significantly less predictors than GC for all target links. If we draw a relationship between the regression tree and GC in the context of number of relevant predictors selected, we can observe a linear relationship as presented in Figure 4.6.

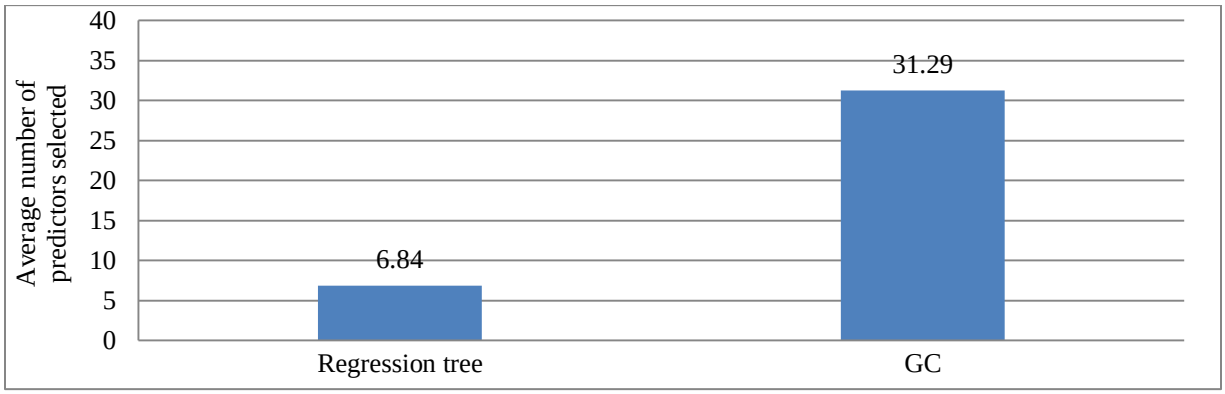


Figure 4.5: Comparison of average number of relevant predictors selected by the regression tree and the Granger causality method.

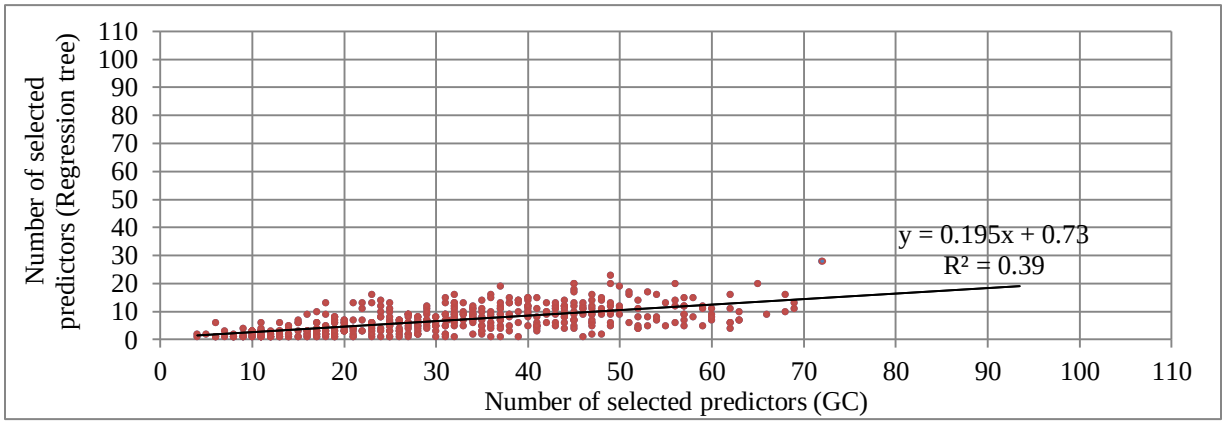
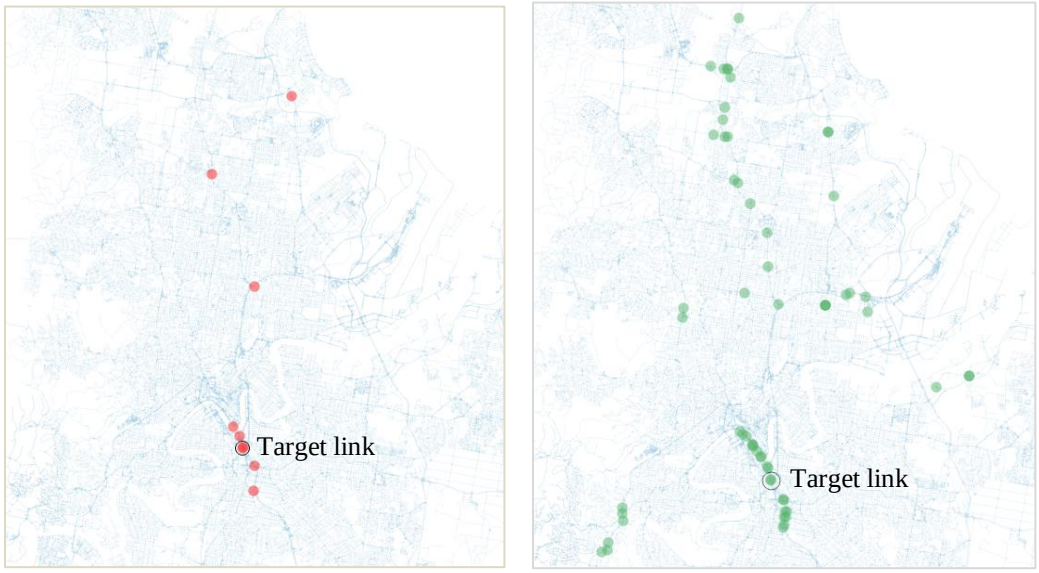


Figure 4.6: Relationship of the number of relevant predictors selected by the regression tree and the Granger causality method.

Figure 4.7 illustrates the location of relevant predictors selected by the regression tree and GC for the target link (Link 80). The selected relevant predictors include nearer neighbour links, distant links of the target link and the target link itself (the past traffic flow). Both methods identify the relevant predictors based on separate statistical approach which determines whether the time series of a predictor link can provide statistically significant information about the future time series of the target link regardless of the distance between the links. Hence, some distant links are selected as the relevant predictor links by the regression tree and GC. It is understandable that nearer neighbour links and the histories of the target link itself have the effects on the future traffic flow of the target link. The selection of some distant links as the relevant predictors indicates the influence of traffic flow of some distant links on the traffic flow of the target link. It can be noted that, compared to GC, the regression tree selects less number of relevant predictors for the target link (Link 80).



a) b)

Figure 4.7: Location of selected relevant predictors for a target link (Link 80 as an example) by a) regression tree b) Granger causality.

4.4.2 Prediction accuracy

This section demonstrates the efficiency of the proposed relevant predictors for each target link as the input variables in short term traffic prediction. The effectiveness is measured in terms of

the prediction accuracy produced by multi-layer feed forward neural network. The relevant predictors identified by vector autoregressive based Granger causality (GC) are also compared with the proposed relevant predictors selected by the regression tree.

The accuracies of the prediction models based on the relevant predictors are assessed by RMSE and MAE. Figure 4.8 compares the average prediction accuracy of the prediction models built by relevant predictors of the regression tree and GC. It can be noticed that the average RMSE and MAE values of the prediction models based on the relevant predictors of regression tree are higher than those of GC. However, the differences are not significantly high, which range approximately 3-4 veh/hr. Furthermore, if we draw a relationship between the precisions of the prediction models based on the regression tree and GC for all target links, we can observe a linear relationship as presented in Figure 4.9. It is evident that only few values are fluctuated from the linear regression line which indicates that the difference of accuracies provided by the regression tree and GC are not significantly higher. Figure 4.5 shows that the average number of relevant predictors selected by the regression tree is approximately one-fourth of the average number of relevant predictors selected by GC. It is understandable that the prediction accuracy increases with the addition of the relevant predictors in the prediction model. Since GC selects significantly higher number of relevant predictors almost all cases, the average prediction accuracy provided by GC is higher than the prediction accuracy provided by the regression tree. Although GC selects the significantly higher number of relevant predictors, it cannot supply significantly higher prediction accuracy compared to the regression tree. This phenomenon indicates the efficiency of the parsimonious set of relevant predictors selected by the regression tree in traffic prediction.

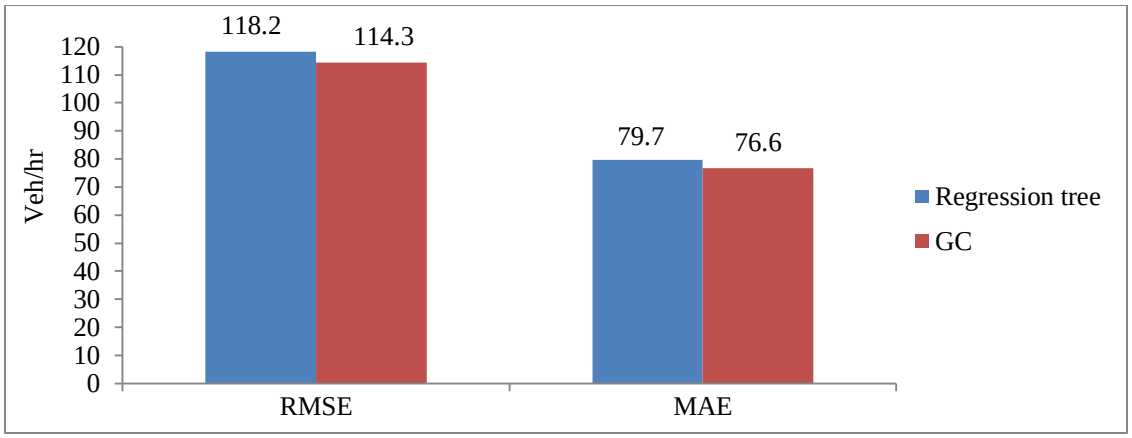
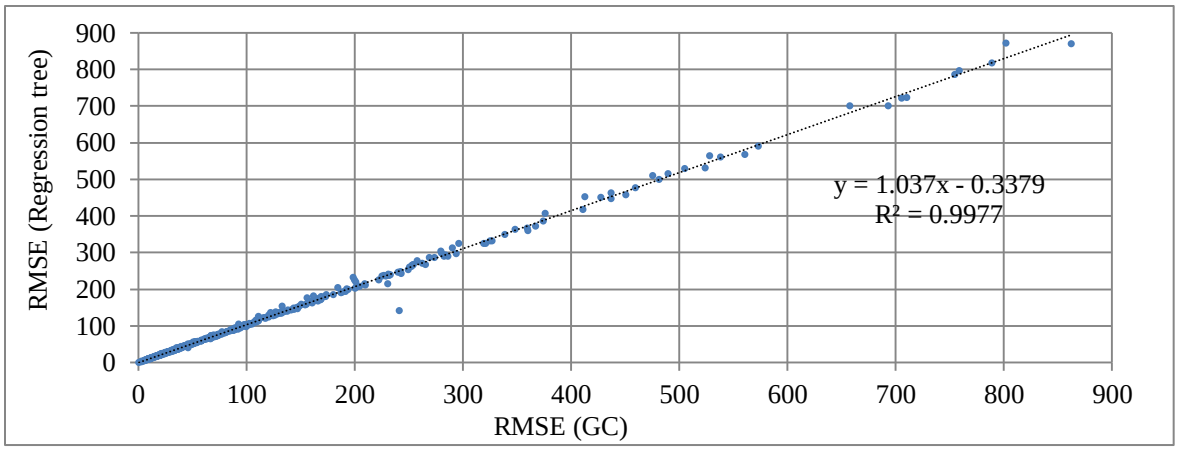
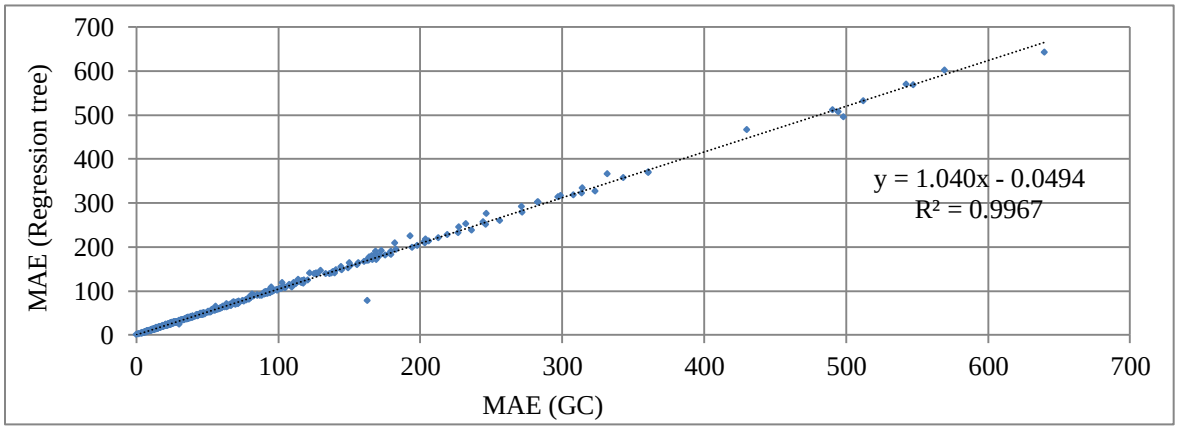


Figure 4.8: Comparison of the prediction accuracies of the neural network based on the predictors selected by the regression tree and the Granger causality.



a) RMSE



b) MAE

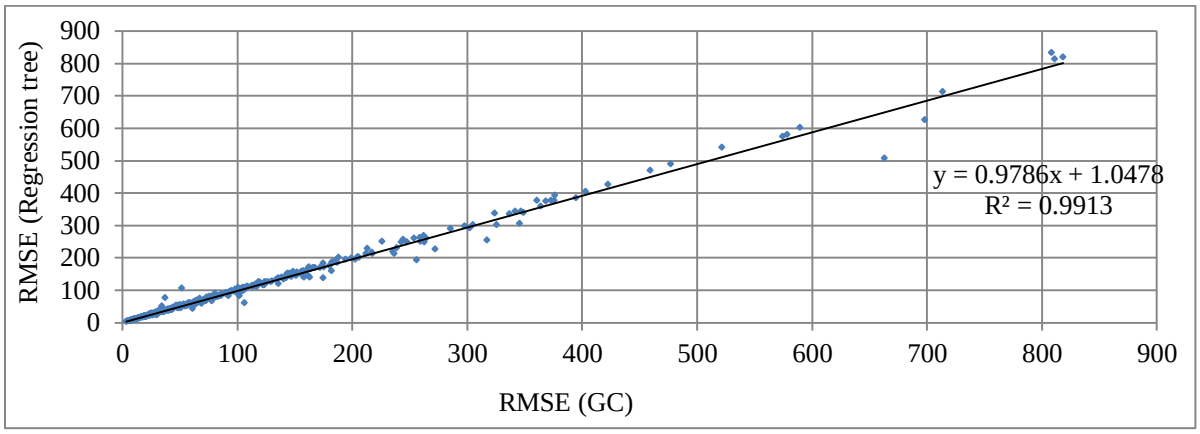
Figure 4.9: Relationship of the prediction accuracies of the neural network based on the predictors selected by the regression tree and the Granger causality.

The regression tree and GC select the past traffic flow of the target link as a relevant predictor for all the 479 target links. If the relevant predictors selected by GC for each target link are ranked according to their GC strength (Bressler and Seth, 2011), the past of the target link is found to be the top ranked predictor for almost all the 479 target links. Thus, the past traffic flow of target link acts as a powerful predictor and it has significant contribution to accurately predicting the future traffic flow of the target link. To evaluate the effectiveness of the remaining relevant predictors selected by the regression tree and GC for each target link, the short term traffic prediction models are built using only the remaining relevant predictors as the input variables.

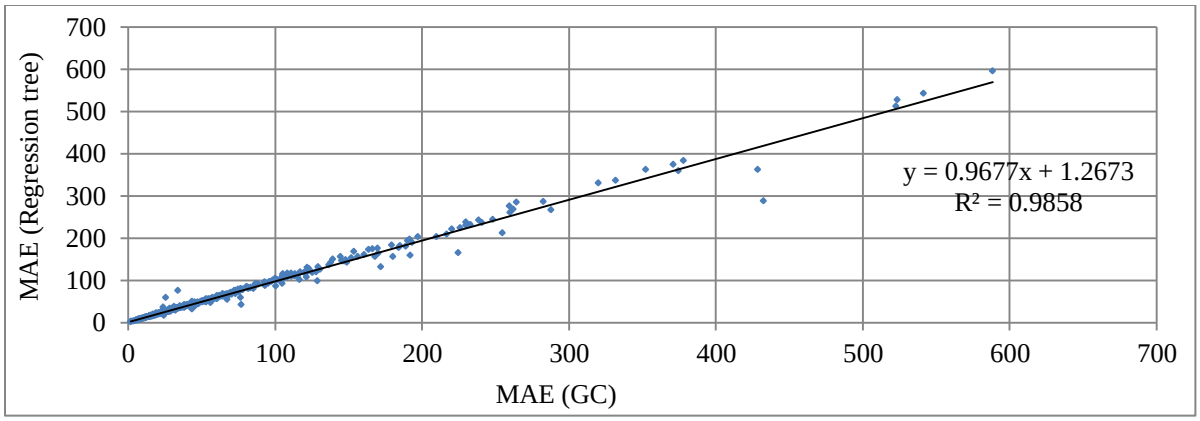
Figure 4.10(a) and Figure 4.10(b) compare the accuracies of the prediction models based on the remaining relevant predictors (i.e. relevant predictors excluding the past of target link) that are identified by the regression tree and GC. In this comparison, we consider the equal number of predictors in GC and the regression tree. Since the number of relevant predictors by the regression tree is less than that of GC in all the 479 target links, the equal number of relevant predictors in both methods can be achieved by reducing to the number of relevant predictors in GC. For this purpose, the selected relevant predictors in GC are ranked according to their GC strength for each of the target link, and then the same number of higher ranked predictors as in the regression tree is taken. Figure 4.10(a) and Figure 4.10(b) show that the relationship between the accuracy of the prediction model based on the remaining relevant predictors by regression tree and GC is linear. Compared to Figure 4.9, more scatter points can be observed in Figure 4.10. This indicates that these two methods produce significantly different predicted values for some target links. Figure 4.11(a) and Figure 4.11(b) illustrate the differences of RMSE and MAE between the prediction models based on the equal number of relevant predictors of GC and the regression tree. The positive values in these figures indicate higher prediction error produced by GC whereas the negative values indicate higher prediction error produced by the regression tree. It is noted that in comparison with regression tree, GC produces significantly higher prediction errors in predicting traffic flow of some target links.

Overall, the outcomes of the short term traffic prediction models reveal that the regression tree is an effective tool to identify more relevant set of input predictors for traffic prediction than GC. As mentioned, the aim of the study is to propose a nonlinear method which can systematically

identify the parsimonious set of predictors that helps to improve the prediction accuracy in short term traffic prediction. The proposed regression tree method allows us to achieve the objective of the study.

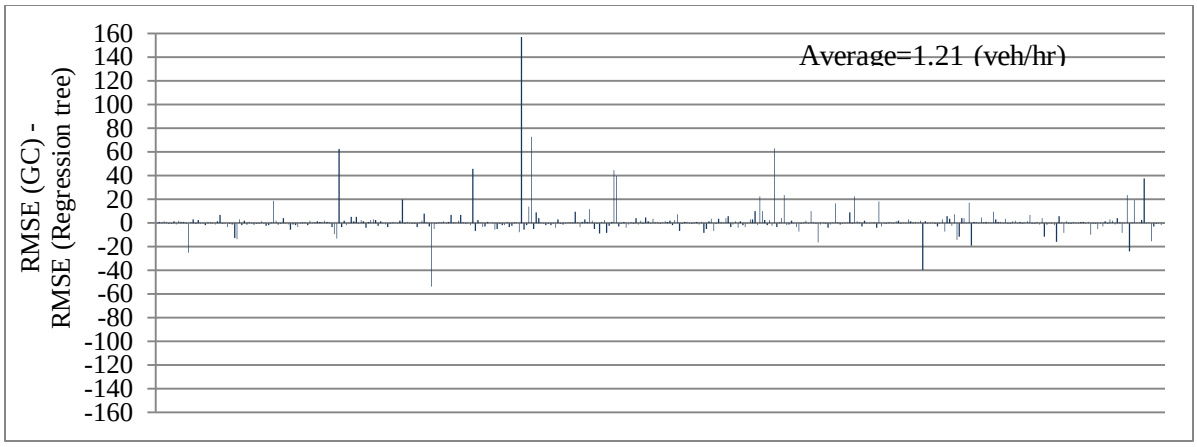


a) RMSE

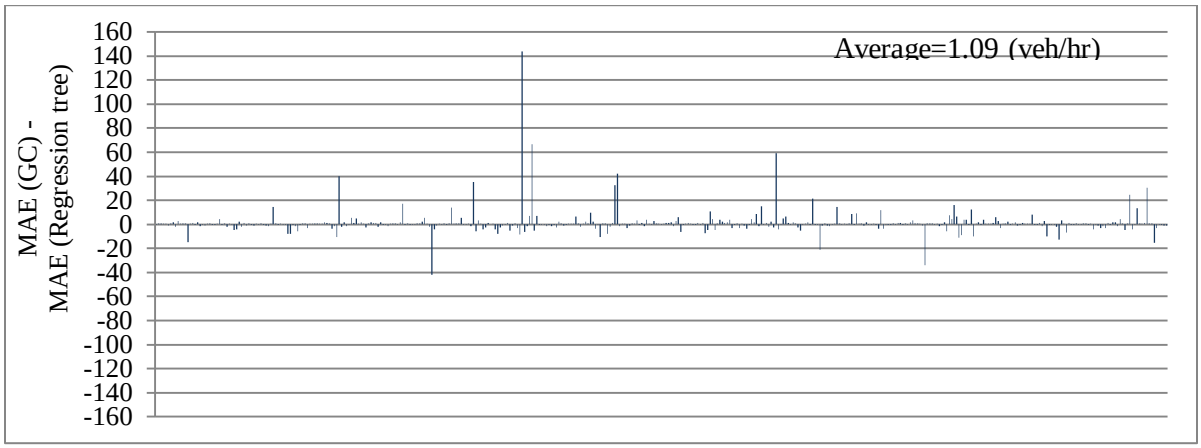


b) MAE

Figure 4.10: Relationship of the prediction accuracies of the neural network based on the relevant predictors except the past of the target link.



a) RMSE



b) MAE

Figure 4.11: Comparison of the prediction accuracies of the neural network based on the relevant predictors set except the past of the target link.

4.5 Conclusion

The study evaluates the spatial dependency of traffic flow among the links in a large-scale road network by exploring the underlying relationship (linear or nonlinear) between the given target link and the predictor links. We propose the regression tree method to select the input variables for short-term traffic prediction models by incorporating the knowledge of the spatial dependency structure of road links in the network. Our proposed method systematically identifies a set of relevant links that have significant statistical influence to the target link. The relevant predictor links selected by the regression tree include neighbour links and some distant links of the target link. The set of relevant links selected by the regression tree and a linear method based on Granger causality is compared in terms of variable reduction and prediction

accuracy for traffic prediction. Compared to the linear method, the regression tree is more efficient in producing a parsimonious set of relevant predictor links for the target links in the network. The prediction accuracy results demonstrate that the regression tree selects more relevant set of predictor links for the target links than the linear method based on Granger causality. Therefore, the regression tree is a powerful variable selection method that minimises the computational cost and maintains the prediction performance of traffic prediction model for a road network. The proposed method can be a valuable tool for traffic management in the urban large-scale road network.

This study implements the proposed regression tree method using the traffic states of the road links over the whole day period. Future research may explore the effectiveness of the regression tree method in peak and off-peak period traffic of a large-scale urban road network.

Chapter 5

CONCLUSIONS

This chapter summarises the results and findings of this research as well as discusses the possible future research directions.

5.1 Research summary and contributions

This study proposes the methods to identify the parsimonious set of relevant predictors as the input predictor variables for short term traffic prediction of the individual target link or the whole network. To select a set of relevant predictors for individual target link, a bivariate linear method is implemented by using pairwise Granger causality test, and to select a common set of important predictors for all the target links in the network, a multivariate linear method is applied by using vector auto regressive Granger causality test and elastic net regularisation. Finally, considering the real traffic condition, a nonlinear method is implemented via regression tree to identify a reduced set of relevant predictors for individual target link. The relevant predictor set identified by each of the methods are tested as the input variables for the traffic prediction model, namely, time series regression and neural network. It is proven from the results that the proposed methods are efficient in reducing the computation complexity of the traffic prediction models and ensuring the higher prediction accuracies.

The summary and outcomes of the thesis are described below:

- The proposed methods are efficient in removing a significant number of irrelevant predictor links and at the same time providing a parsimonious set of relevant predictor links for each target link in the road network. The computational challenges of traffic prediction in a large-scale road network that includes hundreds of links can be reduced by using the parsimonious set of relevant predictor links as the input predictor variables. The results of the study show that the set of the relevant predictor links for each target link identified by the proposed methods reduces the computational complexity in developing

short term traffic prediction model. The proposed variable selection techniques that incorporate the knowledge of the spatial dependency structure of the road network can therefore be an effective tool for traffic management and control scheme in a large-scale road network.

- The proposed methods reveal that the relevant predictor links for a target link can be found across the entire network whereas the existing methods mainly focus on neighbouring links (when not considering only the past traffic states of the target link or simply the upstream and downstream links). For each target link, the relevant links identified by the proposed methods include a number of neighbour links and distant links of the target link. It indicates that traffic flow of a target link not only depends on its past traffic flow but also depends on the past traffic flow of the neighbour and distant locations. The implementation of the relevant set of predictors in short term traffic prediction shows that it is effective to use a spatially diverse set of relevant predictor links as the input predictor variables.

- The relevant predictor links selected by the proposed methods are the potential predictor variables for ensuring higher prediction accuracy of the linear and nonlinear short term traffic prediction model such as time-series regression and artificial neural network. The prediction results in the case-studies show that the traffic prediction models based on the relevant predictor links outperform the traffic prediction models that use only upstream-downstream links, neighbourhood links or randomly selected links. Furthermore, explicitly considering the input variable selection as a pre-processing step of traffic prediction provides better prediction accuracy than some popular traffic prediction models (e.g. ARIMA).

- Our proposed bivariate linear method quantifies the dependence of a target link to the other links in the network by Granger causal-strength and prepares a hierarchy of the influence of the predictor links. By applying the 90^{th} percentile threshold of Granger causal-strength, a parsimonious set of the most significant predictor links for each target link is identified. This method of selecting the most significant links leads to the higher

prediction accuracy regardless of the target link and the prediction model. It suggests Granger causal-strength is an effective criterion to prepare a ranking of the predictor links and to select the most significant predictor links for each target link.

- This study focuses on the traffic states of road links over the whole day period to implement the proposed methods of selecting the most significant predictor links for each target link. The application of the proposed bivariate linear method on the peak and off-peak period traffic state data, namely, morning peak period, daytime off-peak period, afternoon peak period and nigh time off-peak period demonstrates that the selection of the set of most significant predictor links for each target link varies with the time period.

- The implementation of the proposed bivariate linear method on speed data also selects a parsimonious set of significant predictor links for each target link; however, a number of selected signifiant predictor links are different from those selected based on traffic flow data. It appears that the selection of the significant predictor links for a target link is dependent on the traffic parameter (i.e. traffic flow and speed) used as the traffic states of the road links.

- This study explores the relationship among the distance of a predictor link from the target link, the optimal time lag of the predictor link and the Granger causal strength of the predictor link. The distance from the target link does not seem to have a relation with the optimal time lag of a predictor link but appears related to the causal-strength of the predictor link since, at least for freeway links, upstream and downstream links show higher values of Granger causal-strength.

- This study demonstrates the advantages of considering the importance of a road link as a measure for selecting the most relevant predictors for the target links in a large-scale network. The proposed approach utilises the spatial relation among the road links in the network to determine the importance of each road link as a predictor and provides a ranking of important predictor links in the network. While existing methods of predictor variable selection identify a distinct set of predictors for each target link, the proposed

approach detects a common set of the most important links that acts as the fixed input predictor variables for any target link in the network. The efficacy of selecting the most important set of links as the predictor variables for short term traffic prediction is shown in context of dimensionality reduction and prediction accuracy. The outcomes of this study can help traffic authorities identify the most critical locations to assign traffic sensors to monitor and predict the network-wide traffic states, given a limited budget and resources. The knowledge of the spatial dependency structure in the network can also help traffic authorities reliably monitor traffic parameters even when there are malfunctioning sensors in the target location by combining the information from its relevant predictor links and filling the gaps for the missing data.

- The proposed nonlinear method of exploring the underlying spatial relationships among road links in the network is effective in identifying a parsimonious set of relevant predictors for each target link in the network. For each target link, the number of the relevant predictors selected by the nonlinear method is significantly less than that of the linear method. The performance of the selected set of relevant predictor links as the input variables in the short term traffic prediction represents the superiority of the proposed nonlinear method over the linear method.

5.2 Limitations and future research directions

A number of possible future research directions are identified as follows:

- The study selects a large-scale road network as the test bed for implementing the proposed methods of spatial variable selection for short term traffic prediction. On the contrary, the existing literature related to traffic prediction selects a small road network as the test bed. The selection of the periphery of the road network to select spatial variables is still user-defined. Finding an appropriate size or periphery of the road network can be an area of future research.

- This study employs a nonlinear method of selecting spatial variables that considers the conditional dependency of the links in the road network. The bivariate (or pairwise) approach-based nonlinear method was not proposed in this study to identify the relevant predictors for a target link. A further study can be conducted to propose a bivariate approach-based nonlinear method of identifying the relevant predictors for a given link.

- For the selection of the network-wide important predictor links, one of the applied multivariate methods in this study is elastic net regularisation. In this method, the user defined value of tuning parameter α can be of any value between 0 and 1. In this study, α=0.5 is used to apply the lasso and ridge penalties equally. The variation of the tuning parameter α can be explored for values other than the assumed 0.5, thus allowing a variation between the ridge and the lasso extremes.

- The proposed method of this study focuses on a single traffic parameter (e.g. traffic flow) to estimate the traffic states of the road links in the network. The selection of the significant predictor links for a target link by the proposed method depends on the traffic parameter used. However, future research will consider using multiple parameters (e.g. traffic flow and speed) together in the proposed method to identify a more generalised set of significant predictor links for the target link.

- The proposed data-driven method of identifying significant predictor links could be further improved by incorporating some traffic models such as shockwave analysis. Shockwave analysis in traffic flow theory describes how traffic states propagate in space and time along a road, and hence, it can identify the links whose traffic states are correlated with a certain time gap. Therefore, using shockwave analysis with the proposed method for selecting the predictor links in the neighbourhood of the target link can be a future research direction.

REFERENCES

Ahmad, J., Harnhirun, S., 1995. Unit roots and co-integration in estimating causality between exports and economic growth: empirical evidence from the ASEAN countries. *Economics Letters*, 49(3), 329-334.

Ahmed, M. S., Cook, A. R., 1979. Analysis of freeway traffic time-series data by using Box-Jenkins techniques. *Transportation Research Record*, 722, 1-9.

Bahadori, M. T., Liu, Y., 2012. On causality inference in time series. In *2012 AAAI Fall Symposium Series*.

Barnett, L., Seth, A. K., 2014. The MVGC multivariate Granger causality toolbox: a new approach to Granger causal inference. *Journal of Neuroscience Methods*, 223, 50-68.

Bernasconi, C., Konig, P., 1999. On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings. *Biological cybernetics*, 81(3), 199-210.

Bishop, C.M., 1995. Neural networks for pattern recognition. *Oxford University Press.*

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., 1984. Classification and regression trees. Wadsworth & Brooks. *Cole Statistics/Probability Series*.

Bressler, S.L., Seth, A.K., 2011. Wiener–Granger causality: a well-established methodology. *Neuroimage*, 58(2), 323-329.

Chandra, S. R., Al-Deek, H., 2009. Predictions of freeway traffic speeds and volumes using vector autoregressive models. *Journal of Intelligent Transportation Systems*, 13(2), 53-72.

Chen, C., Wang, Y., Li, L., Hu, J., Zhang, Z., 2012. The retrieval of intra-day trend and its influence on traffic prediction. *Transportation research part C: emerging technologies*, 22, 103-118.

Chu, X., Wu, C., Qiu, J., 2016. A nonlinear Granger causality test between stock returns and investor sentiment for Chinese stock market: a wavelet-based approach. *Applied Economics*, *48*(21), 1915-1924.

Clark, S., 2003. Traffic prediction using multivariate nonparametric regression. *Journal of Transportation Engineering*, 129 (2), 161-167.

Cottrell, A., Lucchetti, R., 2016. GNU regression, econometrics and time-series library. Retrieved from: http://gretl.sourceforge.net. Accessed: June 8, 2016.

Dhamala, M., Rangarajan, G., Ding, M., 2008. Analyzing information flow in brain networks with nonparametric Granger causality. *NeuroImage*, 41(2), 354-362.

Demirbas, S., 1999. Cointegration Analysis-Causality Testing and Wagner's Law: The Case of Turkey, 1950-1990. *Discussion Papers in Economics*, Department of Economics, University of Leicester, United Kingdom.

Dickey, D.A., Fuller, W.A., 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a), 427-431.

Ding, M., Chen, Y., Bressler, S. L., 2006. Granger causality: basic theory and application to neuroscience. *Handbook of time series analysis: recent theoretical developments and applications*, 437.

Durbin, J., Watson, G.S., 1971. Testing for serial correlation in least squares regression III. *Biometrika*, 58(1), 1-19.

Ermagun, A., Levinson, D., 2018. Spatiotemporal traffic forecasting: review and proposed directions. *Transport Reviews*, 38(6), 786-814.

Fonarow, G. C., Adams, K. F., Abraham, W. T., Yancy, C. W., Boscardin, W. J. and ADHERE Scientific Advisory Committee, 2005. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. *Jama*, 293(5), 572-580.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalised linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1-22.

Geweke, J., 1982. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77(378), 304-313.

Geweke, J.F., 1984. Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association*, 79(388), 907-915.

Goebel, R., Roebroeck, A., Kim, D. S., Formisano, E., 2003. Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping. *Magnetic resonance imaging*, 21(10), 1251-1261.

Gosh, B., Basu, B., O'Mahony, M., 2007. Bayesian time-series model for short-term traffic flow forecasting. *Journal of Transportation Engineering*, 133(3), 180-189.

Goves, C., North, R., Johnston, R., Fletcher, G., 2016. Short term traffic prediction on the UK motorway network using neural networks. *Transportation Research Procedia*, 13, 184-195.

Granger, C. W., 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37(3), 424-438.

Granger, C. W., 1980. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and Control*, 2, 329-352.

Günther, F., Fritsch, S., 2010. Neuralnet: Training of neural networks. *The R journal*, 2(1), 30-38.

Hamilton, J. P., Chen, G., Thomason, M. E., Schwartz, M. E., Gotlib, I. H., 2011. Investigating neural primacy in Major Depressive Disorder: multivariate Granger causality analysis of resting-state fMRI time-series data. *Molecular psychiatry*, 16(7), 763.

Hasan, M. M., Kim, J., 2016, Analysing functional connectivity and causal dependence in road traffic networks with Granger causality. *In 38th Australasian Transport Research Forum (ATRF) Proceedings*, Melbourne, Australia, 16-18 November.

Hasan, M. M., Kim, J., Prato, C., 2017. Spatial variable selection methods for network-wide short-term traffic prediction. *In 39th Australasian Transport Research Forum (ATRF) Proceedings*, Auckland, New Zealand, 27-29 November.

Hastie, T., Qian, J., 2014. Glmnet vignette. Retrieved from: http://www.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf. Accessed: September 20, 2016.

Hobeika, A. G., Kim, C. K., 1994. Traffic-flow-prediction systems based on upstream traffic. *Proceedings of the Vehicle Navigation and Information Systems Conference, IEEE,* 345-350.

Hoerl, A., Kennard, R., 1988. Ridge regression. *Encyclopedia of Statistical Sciences*, 8, 129-136.

Hou, Y., Edara, P. and Sun, C., 2014. Traffic flow forecasting for urban work zones. *IEEE transactions on intelligent transportation systems*, 16(4), 1761-1770.

Kamarianakis, Y., Prastacos, P., 2003. Forecasting traffic flow conditions in an urban network: comparison of multivariate and univariate approaches. *Transportation Research Record*, 1857, 74-84.

Kamarianakis, Y., Prastacos, P., 2004. Space–time modeling of traffic flow. *Computers & Geosciences*, 31(2), 119-133.

Kamarianakis, Y., Gao, H.O., Prastacos, P., 2010. Characterizing regimes in daily cycles of urban traffic using smooth-transition regressions. *Transportation Research Part C: Emerging Technologies*, 18(5), 821-840.

Karlaftis, M.G., Golias, I., 2002. Effects of road geometry and traffic volumes on rural roadway accident rates. *Accident Analysis & Prevention*, 34(3), 357-365.

Kumar, K., Parida, M., Katiyar, V.K., 2013. Short term traffic flow prediction for a non urban highway using artificial neural network. *Procedia-Social and Behavioral Sciences*, 104, 755-764.

Lawrence, R.L., Wright, A., 2001. Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric engineering and remote sensing*, 67(10), 1137-1142.

Lee, K. J., Carlin, J. B., 2010. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American journal of epidemiology*, 171(5), 624-632.

Lewis, R. J., 2000. An introduction to classification and regression tree (CART) analysis. In *Annual meeting of the society for academic emergency medicine*, San Francisco, California, 14.

Li, L., Su, X., Wang, Y., Lin, Y., Li, Z., Li, Y., 2015. Robust causal dependence mining in big data network and its application to traffic flow predictions. *Transportation Research Part C: Emerging Technologies*, 58, 292-307.

Li, Q., Lin, N., 2010. The Bayesian elastic net. *Bayesian Analysis*, 5(1), 151-170.

Li, Z., Jiang, S., Li, L., Li, Y., 2019. Building sparse models for traffic flow prediction: An empirical comparison between statistical heuristics and geometric heuristics for Bayesian network approaches. *Transportmetrica B: Transport Dynamics*, 7(1), 107-123.

Min, W., Wynter, L., 2011. Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies*, 19(4), 606-616.

Ogutu, J. O., Schulz-Streeck, T., Piepho, H. P., 2012. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *Proceedings of*

the 15th European workshop on QTL mapping and marker assisted selection (QTLMAS), *BMC Proceedings*, 6(2), 10.

Okutani, I., Stephanedes, Y. J., 1984. Dynamic prediction of traffic volume through Kalman filtering theory. *Transportation Research Part B: Methodological*, 18(1), 1-11.

Pavlidis, E.G., Paya, I., Peel, D.A., 2015. Testing for linear and nonlinear Granger causality in the real exchange rate–consumption relation. *Economics Letters*, 132, 13-17.

Pavlyuk, D., 2019. Feature selection and extraction in spatiotemporal traffic forecasting: a systematic literature review. *European Transport Research Review*, 11(1), 6.

Pascale, A., Nicoli, M., 2011. Adaptive Bayesian network for traffic flow prediction. *Statistical Signal Processing Workshop (SSP), IEEE*, 177-180.

Queen, C. M., Albers, C. J., 2009. Intervention and causality: forecasting traffic flows using a dynamic Bayesian network. *Journal of the American Statistical Association*, 104(486), 669-681.

Ramsey, J. B., 1969. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(2), 350-371.

Ramsey, J. B., 1974. Classical model selection through specification error tests. *Frontiers in econometrics*, 1, 13-47.

Seth, A.K., 2010. A MATLAB toolbox for Granger causal connectivity analysis. *Journal of Neuroscience Methods*, 186(2), 262-273.

Seth, A. K., Barrett, A. B., Barnett, L., 2015. Granger causality analysis in neuroscience and neuroimaging. *The Journal of Neuroscience*, 35(8), 3293-3297.

Siddiqui, C., Abdel-Aty, M., Huang, H., 2012. Aggregate nonparametric safety analysis of traffic zones. *Accident Analysis & Prevention*, 45, 317-325.

Smith, B. L., Williams, B. M., Oswald, R. K., 2002. Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies*, 10(4), 303-321.

SPSS Inc., 2008. SPSS Statistics for Windows, version 17.0, Chicago: SPSS Inc.

Stathopoulos, A., Karlaftis, M. G., 2003. A multivariate state space approach for urban traffic flow modeling and prediction. *Transportation Research Part C: Emerging Technologies*, 11(2), 121-135.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., Carpenter, J. R., 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*, 338, b2393.

Sun, S., Zhang, C., Yu, G., 2006. A Bayesian network approach to traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 7(1), 124-132.

Sun, S. and Zhang, C., 2007. The selective random subspace predictor for traffic flo forecasting. *IEEE Transactions on intelligent transportation systems*, 8(2), 367-373.

Thalassinos, I.E., Pintea, M., Raţiu, I. P., 2015. The Recent Financial Crisis and Its Impact on the Performance Indicators of Selected Countries during the Crisis Period: A Reply. *International Journal of Economics and Business Administration*, 3(1), 3-20.

Therneau, T. M., Atkinson, E. J., 1997. An introduction to recursive partitioning using the RPART routines. Retrieved from: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf. Accessed: October 25, 2018.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

Vančo, B. A., 2012 Quantification of causal interactions in complex systems. *Master Thesis.* Comenius University, Bratislava. Retrieved from: http://cogsci.fmph.uniba.sk/~farkas/theses/anton.vanco.dip12.pdf. Accessed: April 19, 2018.

Wang, J., Deng, W., Guo, Y., 2014. New Bayesian combination method for short-term traffic flow forecasting. *Transportation Research Part C: Emerging Technologies*, 43, 79-94.

Xu, Y., Kong, Q.J., Klette, R., Liu, Y., 2014. Accurate and interpretable Bayesian mars for traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 15(6), 2457-2469.

Xu, Y., Chen, H., Kong, Q.J., Zhai, X., Liu, Y., 2016. Urban traffic flow prediction: a spatio-temporal variable selection-based approach. *Journal of Advanced Transportation*, 50(4), 489-506.

Yang, S., Shi, S., Hu, X., Wang, M., 2015. Spatiotemporal context awareness for urban traffic modeling and prediction: sparse representation based variable selection. *PloS one*, 10(10), e0141223.

Yang, S., Wu, J., Du, Y., He, Y., Chen, X., 2017. Ensemble learning for short-Term traffic prediction based on gradient boosting machine. *Journal of Sensors*, 2017.

Yu, L., Li, J., Tang, L., Wang, S., 2015. Linear and nonlinear Granger causality investigation between carbon market and crude oil market: A multi-scale approach. *Energy Economics*, 51, 300-311.

Zhang, C., Ren, J., 2013. GCBN: a hybrid spatio-temporal causal model for traffic analysis and prediction. *International Conference on Web-Age Information Management*, Springer, Berlin, Heidelberg, 265-276.

Zivot, E., Wang, J., 2006. Vector autoregressive models for multivariate time series. *Modeling Financial Time Series with S-PLUS®*, 385-429.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2), 301-320.

Zou, H., Zhang, H. H., 2009. On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4), 1733-1751.